Науково-виробниче підприємство "Ідея"

Бібліотечний форум України

http://www.idea.com.ua/forum/

№4 2005

Влащенко Л., Грищенко Т., Нікітенко О.М. Технологія створення колекції повнотекстових електронних видань у бібліотеках //Бібл. форум України.-2004.- №4.- С.26.

Інтернет-технології в бібліотеках

ТЕХНОЛОГІЯ СТВОРЕННЯ КОЛЕКЦІЇ ПОВНОТЕКСТОВИХ ЕЛЕКТРОННИХ ВИДАНЬ У БІБЛІОТЕКАХ

Людмила Влащенко,
завідувачка відділу автоматизації,

Тамара Грищенко,
директор наукової бібліотеки,

Олександр Нікітенко,
кандидат технічних наук, доцент Харківський національний університет радіоелектроніки

Наведено порівняння графічних форматів за обсягом, подано короткий опис графічного формату DjVu. Обґрунтовано застосування цього формату для створення повнотекстових електронних колекцій. Спираючись на досвід створення повнотекстових електронних документів формату DjVu, наведено технологію виготовлення таких документів із застосуванням відповідних програмних пакетів.

Вступ

Одним з найважливіших показників зміни способу життя у XXI столітті є розвиток і використання прогресивних інформаційних технологій в усіх сферах суспільного життя та діяльності, рівень виробництва та споживання суспільством інформаційних продуктів та послуг. У цих умовах важливою справою є формування нової телекомунікаційної культури українського суспільства та вирішення проблеми якісно нової освіти.

Нові інформаційні технології призвели до принципових змін й у системі бібліотечного сервісу. Зникла необхідність багатьох інформаційних форм бібліографічної діяльності й розвинулися зовсім інші форми пошуку та зберігання навчальної та наукової інформації. На жаль, у більшості наукових бібліотек університетів мережеві комп'ютерні технології роблять тільки перші кроки.

Наразі все більшої популярності набувають колекції з повнотекстовими електронними виданнями, при цьому актуальною є проблема створення та доставки користувачу таких видань в електронному вигляді.

Під повнотекстовими електронними виданнями маємо на увазі будь-який документ в електронному вигляді (аудіо, відео, графічний, текстовий тощо).

Вкрай незадовільний стан щодо фінансування бібліотек на придбання навчальної літератури призводить до того, що вона надходить лише у декількох примірниках, у той час як потреба в ній нараховується сотнями.

Отже, завданням бібліотеки є необхідне забезпечення навчального і наукового процесів університету потрібною літературою.

Для досягнення поставленої мети необхідно вирішити такі завдання:

1) верифікувати джерела надходження видань;

2) визначитися з форматом повнотекстових електронних документів;

3) скласти бібліографічний опис електронних видань;

4) здійснити доставку електронних видань користувачеві.

1. Джерела надходження електронних видань

Перш ніж створити колекції електронних документів, необхідно визначитися, з яких джерел надходитимуть ці документи.

Для забезпечення навчальною літературою з кожної дисципліни, яка викладається в Харківському національному університеті радіоелектроніки, обов'язково видаються методичні вказівки та навчальні посібники. Оскільки для видання

такого роду літератури її готують в електронному вигляді, за наказом ректора всі електронні копії передаються до бібліотеки, а це і є одним із джерел надходження електронних видань.

Другим джерелом надходжень є Інтернет, де зберігається велика кількість електронних документів. Але робота з Інтернетом потребує значних витрат ресурсів, часу і постійного моніторингу як вже відомих сайтів, так і нових.

Третє джерело — кафедри університету. Майже всі кафедри університету мають власні електронні бібліотеки, де зберігаються електронні видання, які відповідають профілю кафедри. З цього джерела також надходять у бібліотеку електронні видання, хоч і у значно меншому обсязі, ніж з перших двох.

Четверте джерело — сканування традиційних видань з подальшою обробкою відсканованого матеріалу.

Таким чином, визначено джерела надходження електронних видань. Надходження з перших трьох джерел вимагають від співробітників бібліотеки мінімальних зусиль: визначення місця розташування матеріалу та складання бібліографічного опису. Четверте джерело бажано розглянути детальніше.

2. Вибір формату

Інформацію в комп'ютері можна зберігати в будь-якому форматі: текстовому, графічному, аудіо, відео. Отже, постає питання, в якому форматі зберігати відскановану інформацію.

Вимоги, які висуваються для таких форматів, це — мінімальний обсяг, зручність роботи, відносно невеликий час обробки після сканування. Зрозуміло, що аудіо - та відеоформати не підходять.

Графічні формати якщо й цікавлять нефахівців, то тільки з погляду розмірів файла.

З текстовою інформацією ситуація зовсім занедбана: добре розуміючи різницю між словом рукописним, друкованим майже не друкується в мережі першоджерела у їхньому реальному форматі. Світові музеї й бібліотеки вже оцифровували більшість рукописів, котрі мають яку-небудь цінність, однак одержані розміри файлів ускладнюють ознайомлення з ними через Інтернет.

З грудня 1997 року в науковій бібліотеці Харківського національного університету радіоелектроніки (ХНУРЕ) було розпочато роботи зі створення власних цифрових ресурсів. При цьому головними вимогами, що висувалися, були простота, зручність та ефективність технологічного процесу виготовлення електронних копій.

Вважаємо, що вищезгадана проблема має розглядатися з урахуванням таких чинників:

- швидкість виготовлення електронної копії;
- розмір створеної копії;
- зручність користування електронними копіями;
- доставка електронної копії користувачеві.

Розглянемо ці чинники детальніше.

Час виготовлення електронної копії поліграфічного видання складається з часу, який витрачається на сканування (якщо вважати, що перетворення твердої копії у м'яку відбувається за допомогою сканера) та обробку відсканованого матеріалу (обробку чернетки електронного документа). Час сканування залишається сталим і залежить тільки від типу сканера, котрим користуються. Отже, при виготовленні електронної копії суттєвим є час обробки відсканованої інформації.

Спираючись на восьмирічний досвід роботи зі сканування у бібліотеці ХНУРЕ, електронні матеріали можна поділити на графічні (формати запису bmp, tif, eps, pcx, djvu, gif, jpg) і текстові (формати запису html, doc, rtf, pdf). Зрозуміло, що найменший час витрачається на обробку графічних форматів bmp, tif, eps, pcx, gif, jpg, бо тут після сканування треба тільки записати відсканований матеріал у відповідний файл. Трохи більше часу витрачається на створення файлів формату DjVu через те, що процес обробки та запису в цьому форматі вимагає більшого часу, ніж попередні. Найдовшою є обробка відсканованого матеріалу в текстових форматах через те, що після сканування та розпізнавання необхідно виправити помилки сканування.

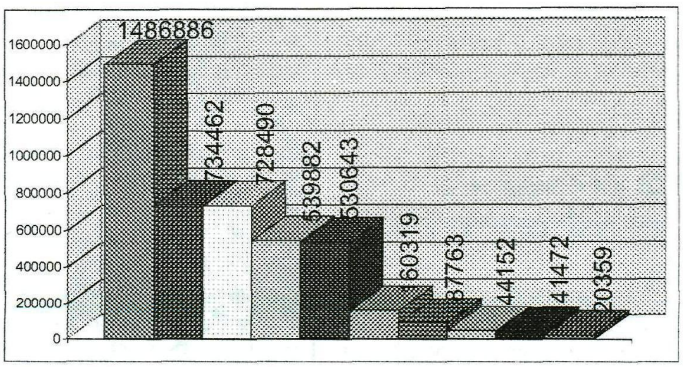
Відомо, що зберігання відсканованого матеріалу потребує певних об'ємів запам'ятовуючих пристроїв, тому при створенні повнотекстових документів треба враховувати й цей чинник. Для з'ясування розмірів форматів було відсканована одну журнальну сторінку формату А4.

При цьому умови сканування були такі: формат А4, розпізнавальна здатність 300 pdi (dots per inch).

На малюнку 1 наведено діаграму розмірів відсканованої сторінки, яка наочно показує, що за розмірами найбільш придатним форматом є DjVu.

Однак всі графічні формати мають суттєвий недолік порівняно з текстовими: кожна сторінка відсканованого тексту міститься в окремому файлі, а це потребує додаткових зусиль для перегляду відсканованого документа.

Наразі активно розробляється новий графічний формат, котрий би задовольняв потреби мережі. Однією з перших закінчених розробок є формат DjVu фірми AT&T, котрий характеризується надзвичайно привабливими характеристиками — застосування нових ефективних алгоритмів архівації дозволило розробникам досягти значного зниження об'ємів файлів. Фірмою розроблено програмні засоби для створення файлів формату DjVu, котрі є безкоштовними для некомерційного використання. Засобом перегляду таких файлів може бути звичайний веб-браузер.



Малюнок 1. Діаграма розмірів відсканованої сторінки

3. Огляд формату DjVu

Новий метод компресії графічних зображень, названий DjVu, створений спеціально для одержання високоякісних копій відсканованих кольорових документів з високим ступенем стиснення, що дозволяє потім швидку передачу образу документа через мережу Інтернет в умовах ліній з низькою швидкістю передачі даних (це особливо актуально для телекомунікаційних каналів країн СНД), відтворювати візуальну копію документа, включаючи колір, шрифт, малюнки і текстуру паперу. Метод DjVu розчленовує зображення на текст і малюнок. Для тексту необхідний більш високий ступінь роздільної здатності, на відміну від малюнків та фону зображення. Малюнок же, звичайно, більш однорідний за своїм складом, тому може бути закодований з меншим ступенем роздільної здатності. Потім застосовується новий метод для максимізації ступеня стиснення: дворівневий передній план зображення кодується за схемою AT&T у новий JBIG2 факс-стандарт, а для фону і малюнків використовується новий метод хвильового стиснення IW44. Обидва методи використовують новий адаптивний арифметичний кодер, названий Z-кодером. Звичайна кольорова сторінка журналу формату A4 із роздільною здатністю 300 dpi може стискатись до розмірів 40–60 Кб, що приблизно в 5–10 разів менше, ніж зображення у форматі JPEG за однакових рівнів ступеня стиснення (див. діаграму). Версія декодера зображення DjVu реалізована у вигляді програмного додатку для всіх популярних веб-браузерів Internet Explorer, Netscape, Opera [1, 2].

Стиснення повнокольорової інформації про документ формату A4 до розміру середньої вебсторінки (46 Кб за даними на 1999 рік) теоретично можливий. Враховуючи зростаючу суспільну потребу в доступі до «оригіналів», здається дивним, що стандарт на графічні зображення такого призначення формується тільки сьогодні [2].

Формат DjVu — перший крок до «кольорового факсу» і його зорієнтовано на передачу, перегляд у мережі й роздрукування переважно текстових документів, для яких важливе значення має не тільки зміст, але й форма: колір та фактура пергаменту, відірваний куточок й сліди від складання вчетверо, ляпка після підпису й круглий відбиток від винної пляшки поруч з печаткою. Архіви всього світу накопичили величезну кількість історичних паперів з неповторним колоритом такого чину.

Існуючі компактні формати JPG, GIF, факс-стандарт CCITT та JBIG забезпечують достатню компресію, однак вузькоспеціалізовані: або на фотографіях, або на чорно-білій графіці й тексті. Тому змішані зображення таких форматів важко прочитати. Розробники DjVu врахували негативний досвід створення «універсального солдата», і новостворений продукт складається з трьох форматів «в одному наборі». Розділ «обов'язків» всередині DjVu базується на простих спостереженнях та фактах.

Текст та інші контрастні малюнки зручно читати при скануванні з розпізнанням не меншим за 300 dpi.

Навпаки, незначна розмитість фонової графіки навіть поліпшує сприйняття тексту. Тому фон без втрат для загального зображення зберігається з розпізнанням 100 dpi в окремому шарі («background»).

Основна проблема — відокремити текст від фону, особливо якщо це кольоровий текст. Здебільшого колір тексту в документах практично однаковий у межах одного знаку. Це дозволяє зберігати кольорову інформацію про текст з розпізнанням всього 25 dpi (шар «foreground») (див. табл. 1).

Шар	Пояснення	Розпізнання, dpi	Глибина кольору, bits/pix
Mask	Монохромна маска-трафарет	300	1
Background	Кольоровий фон	100	24
Foreground	Кольори маски	25	24

Таблиця 1. Розподіл у файлі формату DjVu

Розподіл зображення на текст і фон (формування шару-маски) базується на так званій мультимасштабній кластеризації. Зображення розбивається на різнорозмірні вкладені сітки, у кожній комірці котрих відбувається розпізнавання текстових та фонових кольорів за максимальними піками на гістограмі. Відокремивши текст від фону в найкрупнішій сітці, алгоритм переходить до уточнення на базі

даних з сіток меншого розміру.

У DjVu для стиснення фону, маски та кольорової інформації про маску застосовують різні алгоритми. Фон стискається вейвлет-алгоритмом IW44 (4x4 wavelets), шар-маска, котра не містить кольорової інформації, стискається методом JB2, що є аналогічним до того, який застосовується у факсах. Кольорова інформація про текст також кодується IW44, але попередньо зменшується до 25 dpi. Формат IW44 є дуже близьким до нового стандарту JPEG2000, але менш вимогливим до системних ресурсів при декомпресії зображення під час перегляду.

Новий формат має широке застосування: он-лайнові книжкові магазини, картографічна інформація і навіть е-хіромантія, де надіслана поштою фотографія долоні обробляється подібним чином.

Наразі цей формат вже здобув широке застосування в бібліотеках різного профілю, у релігійних організаціях, радіоаматорських колах. Використовуючи формат DjVu, видаються он-лайнові математичні журнали.

Перехід до DjVu з його чітким текстом, на думку експертів, почнеться з сайтів ЗМІ, котрі копіюють свої паперові видання.

Формат DjVu дозволяє швидко переглянути матеріал у відкритому вигляді, а вже потім вирішити, чи варто його зберігати.

Якщо врахувати, що сторінка чорно-білої графіки з текстом має обсяг у форматі DjVu біля 30 Кб, а у кольорі біля 60 Кб, то стає зрозумілою економія часу і грошей.

Досить об'єктивна оцінка якості в порівнянні з уже відомими форматами показала, що незначне погіршення якості на кольорових зображеннях повністю компенсується ступенем стиснення, котрий сягає сотень і тисяч разів, а на чорно-білих зображеннях погіршення практично не видно. Можливі конкуренти у вигляді tiff, gif, jpg сильно програють в обсязі. До речі, популярний jpg зовсім непридатний для чорно-білих відсканованих зображень. Саме на чорно-білих схемах і текстах, а поліграфічні видання здебільшого такими і є, перевага DjVu є колосальною.

Таким чином, вибір зупинився на форматі DjVu, котрий розробляється американською компанією LizardTech. Формат DjVu реально дозволяє здійснити надто велике стиснення зображень високої якості. Основні переваги цього формату: доступ до цифрової колекції по мережі Internet/Intranet з використанням стандартного програмного забезпечення (необхідно лише встановити додатковий модуль для браузера); висока якість та малий об'єм зображень будь-яких видів (20–30 Кб для чорно-білого зображення формату А4 з розпізнанням 300 dpi; 80–100 Кб для такого ж повнокольорового зображення); повне збереження виду видання; орієнтація на середовище Internet/Intranet і простота забезпечення навігації всередині публікації [2]. Недоліком такого рішення є те, що сторінка у DjVu-форматі є зображенням, і не дозволяє використовувати будь-який пошук. Власне, сам формат DjVu дозволяє зберігати текстову інформацію після процедури OCR (оптичного розпізнавання символів).

Зрозуміло, що бібліотеки мають обмаль коштів, котрих ледве вистачає на комплектування, тому бажано мати програмні пакети за мінімальною ціною. На щастя, в мережі Інтернет існують безкоштовні пакети, котрі дозволяють відтворювати повноцінні зображення у форматі DjVu, включаючи й багатосторінкові файли, перелік таких пакетів наведено в табл. 2.

4. Створення файлів у форматі DjVu

Маючи набір програмних продуктів, що наведені в табл. 2, користувач може створювати файли формату DjVu будь-якої конфігурації.

Алгоритм технології створення файлу формату DjVu, який упроваджений у науковій бібліотеці ХНУРЕ, наведено на мал. 2.

Сканування твердої копії відбувається за допомогою будь-якого програмного пакету, який дозволяє здійснювати сканування з подальшим зберіганням відсканованого матеріалу в будь-якому графічному форматі. Надалі за допомогою будь-якого графічного редактора усуваються вади сканування.

Після сканування та обробки всього необхідного матеріалу здійснюється створення багатосторінкових файлів. Це може бути як один багатосторінковий файл, котрий містить усі відскановані файли (операція bundle), так і створений додатково індексний файл (операція join). Обидві ці операції можна виконати, використовуючи програму DjVuSolo 3.1, котра дозволяє здійснювати ці операції у віконному режимі.

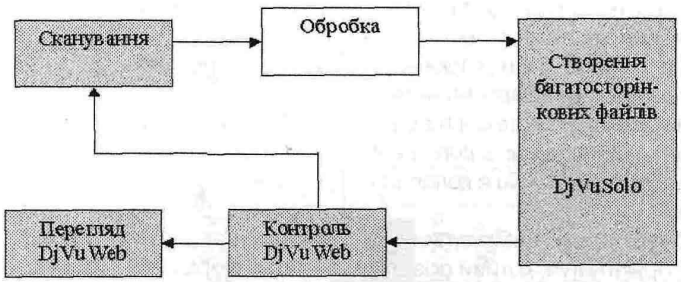
Процедура створення багатосторінкового файла DjVu з окремих графічних файлів полягає у наступному.

По-перше, у вікно програми DjVuSolo вставляють першу сторінку.

По-друге, додають усі інші файли в режимі "Open" таким

Назва	Обсяг, Кб	Призначення
DjVuShop 2.0	1613	Створення, редагування та перегляд односторінкових документів формату DjVu.
DjVuSolo 3.1	2228	Створення, редагування та перегляд одно й багатосторінкових документів формату DjVu.
DjVuMulti 3.0	2099	Утиліти командного рядка для створення багатосторінкових документів формату DjVu.
DjVuSDK2	1907	Утиліти командного рядка для перетворення односторінкових документів формату DjVu.
DjVuWebBrowser	1805	Перегляд документів формату DjVu безпосередньо у браузерах.

Таблиця 2. Програмні продукти для створення файлів формату DjVu



Малюнок 2. Технологія створення файла формату DjVu

чином: позначаємо останню сторінку з групи файлів, які необхідно додати, і першу сторінку за допомогою клавіші Shift — таким чином додається група файлів у необхідній послідовності до багатосторінкового файла.

Через те що створення багатосторінкового файла є дуже громіздкою і відповідальною роботою, необхідним є візуальний контроль створеного файла, котрий здійснюється програмою DjVuSolo 3.1, його переглядом за допомогою будь-якого браузера, до котрого додано програму DjVuWebBrowser.

Після проведення всіх операцій створений багатосторінковий файл можна пропонувати для використання.

Таким чином обґрунтовано вибір формату DjVu для створення повнотекстових копій друкованих видань та описано процедуру створення таких файлів.

5. Бібліографічний опис

Бібліографічний опис електронних документів має свої особливості та залежить від вигляду документа. Електронний документ може бути суто електронним і не мати друкованої копії в бібліотеці, або таким, що має друковану копію у бібліотеці.

Бібліографічний опис суто електронного документа здійснюється за ГОСТом 7.82–2000.

Через те що наукова бібліотека Харківського національного університету радіоелектроніки працює з системою «УФД/Бібліотека», крім власне бібліографічного опису для електронного документа першого вигляду, у полі «Тип літератури» обов'язково вказується значення «Електронний ресурс», для інших типів — залежно від типу літератури.

Після заповнення всіх необхідних полів у формі «Бібліографічний опис документа» необхідно виконати одну дуже важливу в цьому випадку операцію: додати до бібліографічного опису місце збереження електронного документа. У системі «УФД/Бібліотека» це здійснюється простим заповненням форми «Електронна копія», де вказується адреса місця збереження електронного документа.

Таким чином здійснюється бібліографічний опис електронного документа з подальшою можливістю роботи з цим документом безпосередньо з електронного каталогу.

6. Електронна доставка документів

Ще одна проблема: електронна доставка документів читачеві. Вона є своєрідним «мостиком» між традиційними бібліотечними та новими інформаційними технологіями, які активніше входять у наше життя. Наразі розпочався період масового впровадження електронної доставки документів (ЕДД) читачеві у практику роботи бібліотек.

Українські служби ЕДД розпочали створюватися зовсім нещодавно. Для реалізації основних технологічних процесів мережевої взаємодії використовуються програми та протоколи мережі Інтернет. Найчастіше електронні копії доставляються зовнішнім замовникам за допомогою електронної пошти. Для читачів бібліотеки Харківського національного університету радіоелектроніки електронні копії розташовуються у вигляді файла або набору файлів на FTP-сервері, а доступ до таких файлів здійснюється через електронний каталог бібліотечно-інформаційної системи «УФД/Бібліотека».

Висновки

Таким чином, у бібліотеці Харківського національного університету радіоелектроніки створюються повнотекстові електронні документи на базі формату DjVu як найбільш придатного. Випробувані технологічні процеси зі створення файлів формату DjVu довели свою простоту, ефективність та економічність. Для нас створення цифрових колекцій стало одним з важливих напрямків діяльності, котрий ми збираємося розвивати й надалі.

Використання формату DjVu можна порадити тим, хто не тільки цінує час, але й вимагає якості. Не варто боятися нового формату: все було колись новим. Краще просто спробувати самому, а потім вирішити, чи вартий новий формат того, щоб його застосовувати.

Список літератури:

1. Негуляев Е. А. Цифровые коллекции в вузовской библиотеке. Концепция и технологические решения / Е. А. Негуляев, Е. А. Охезина // Библиотеки и ассоциации в меняющемся мире: новые технологии и новые формы сотрудничества. Тема 2002 года: Электронные информационные ресурсы и социальная значимость библиотек будущего: Тр. конф. / 9-я междунар. конф. «Крым-2002». — М.: ГПНТБ России, 2002. — Т. 1. — С. 271–274.

2. Волохін О. «Дежа вю» — нові технології для цифрової бібліотеки // Бібліотечна планета. — 2001. — № 3. — С. 9–11.