МІНІСТЕРСТВО АГРАРНОЇ ПОЛІТИКИ УКРАЇНИ

ВІСНИК ХАРКІВСЬКОГО ДЕРЖАВНОГО ТЕХНІЧНОГО УНІВЕРСИТЕТУ СІЛЬСЬКОГО ГОСПОДАРСТВА

Випуск 17

"ПІДВИЩЕННЯ НАДІЙНОСТІ ВІДНОВЛЮЄМИХ ДЕТАЛЕЙ МАШИН"

Харків 2003

ДК 02(063)

СТВОРЕННЯ ПОВНОТЕКСТОВИХ ДОКУМЕНТІВ У ФОРМАТІ DJVU

Влащенко Л.Г., інженер, Нікітенко А.Н., канд. техн. наук.,
Влащенко Г.І., канд. техн. наук
(Харківський державний технічний університет сільського господарства,
Харківський національний університет радиоелектроники)

В статті наведено порівняння графічних форматів за обсягом, подано опис нового графічного формату DjVu. Спираючись на досвід створення повнотекстових електронних документів формату DjVu наведено технологію виготовлення таких документів з застосуванням відповідних програмних пакетів.

Одним з найважливіших показників зміни образу життя у XXI столітті буде розвиток та використання прогресивних інформаційних технологій у всіх сферах соціального 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 технологій призвели до принципових змін й у системі бібліотечного сервісу. Зникла необхідність у багатьох інформаційних формах бібліографічної діяльності й розвинулися зовсім нові форми пошуку та зберігання учбової та наукової інформації. Нажаль у більшості наукових бібліотек університетів мереживі 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 у електронному вигляді.

Тут під повнотекстовими документами мається на увазі будь-який документ у електронному вигляді (аудіо, графічний, текстовий).

Вибір формату

Графічні формати якщо й цікавлять нефахівців, то тільки з точки зору розмірів файлу.

З текстовою інформацією ситуація зовсім занедбана: добре розуміючи різницю між словом рукописним та друкованим, ми майже не друкуємо в мережі першоджерел у їх реальному форматі. Світові музеї й бібліотеки вже оцифрували більшість рукописів, котрі мають яку-небудь цінність, однак розміри файлів, що отримано, не дозволяють ознайомитися з ними через Internet.

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

Ми бачимо, що вище згадана проблема має розглядатися з урахуванням таких чинників:

1. Швидкість виготовлення електронної копії
2. Розмір створеної копії
3. Зручність користування електронними копіями
4. Доставка електронної копії до користувача

Розглянемо ці чинники детальніше.

Час виготовлення електронної копії поліграфічного видання складається з часу, що витрачається на сканування (вважаємо що перетворення твердої копії у м'яку відбувається за допомогою сканера) та обробки відсканованого матеріалу (обробки чернетки електронного документу). 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 чотирирічний досвід роботи зі сканування у бібліотеці ХНУРЕ електронні матеріали можна поділити на графічні (формати запису bmp, tif, eps, pcx, djvu, gif, jpg) та 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 у цьому форматі вимагає більшого часу ніж попередні. Найдовшою є обробка відсканованого матеріалу у 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 відсковано одну сторінку формату А4, яку наведено на рис. 1.

При цьому умови сканування були такі: формат А4, розрізнювальна здатність 300 pdi, режим сканування line art.

На рис. 2 наведено діаграму розмірів відсканованої сторінки. З цього рисунка випливає, що за розмірами найбільш придатним форматом є djvu.

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

Наразі здійснюються активні розробки нового графічного формату, 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 стиску дозволило розробникам досягнути 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 web-броузер. Було здійснено роботи з освоєння технології створення документів формату DjVu.

На рис. 3 показано, что действия обеих сторон негативными альтернативами (–/–) свидетельствуют о том, что с помощью «войн» понять друг друга нельзя. Позитивные действия с обеих сторон приводят к мирному исходу. Варианты альтернатив (–/+) или (+/–) могут привести к мирному варианту согласия, что определяется цепочкой причинно-следственных альтернатив в многоходовом взаимодействии.

Состояние человека во взаимодействии можно интерпретировать графически в виде сочетания степени его активности и уровня настроения (рис. 4).

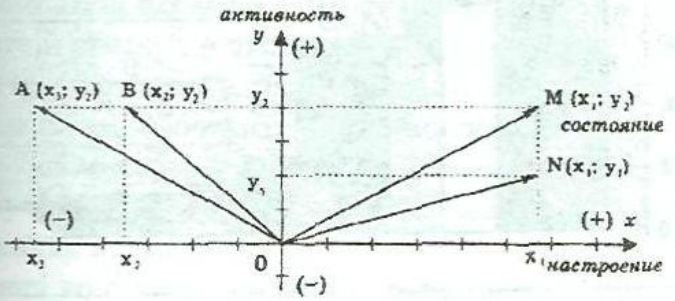

Рис. 4.
Графическая модель оценки состояния партнера

Измерение этих показателей можно производить от среднего, нейтрального (0) уровня. Тогда точка состояния определяется вектором с соответствующими координатами, например $M(x_1, y_2)$. Состояние, определяемое другим вектором $N(x_1, y_1)$, отличается меньшей активностью $Dy = (y_2 - y_1)$. Состояние партнера, определяемое вектором $A(x_3, y_2)$, отличается более скверным настроением, чем состояние партнера, определяемое вектором $B(x_2, y_2)$.

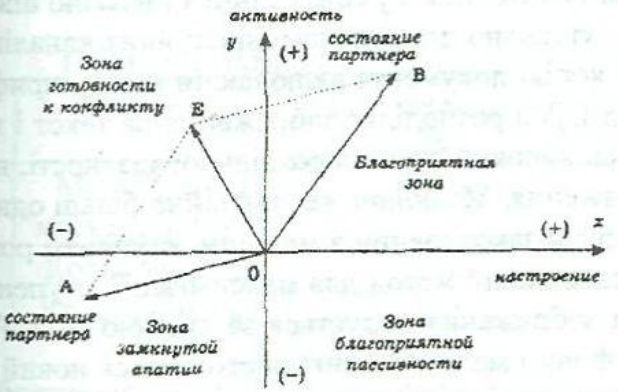

Рис. 5.
Сетевые модели ролевого взаимодействия партнеров

«Менеджмент в России и за рубежом», № 6, 2001

Рис. 1. Приклад сторінки для сканування

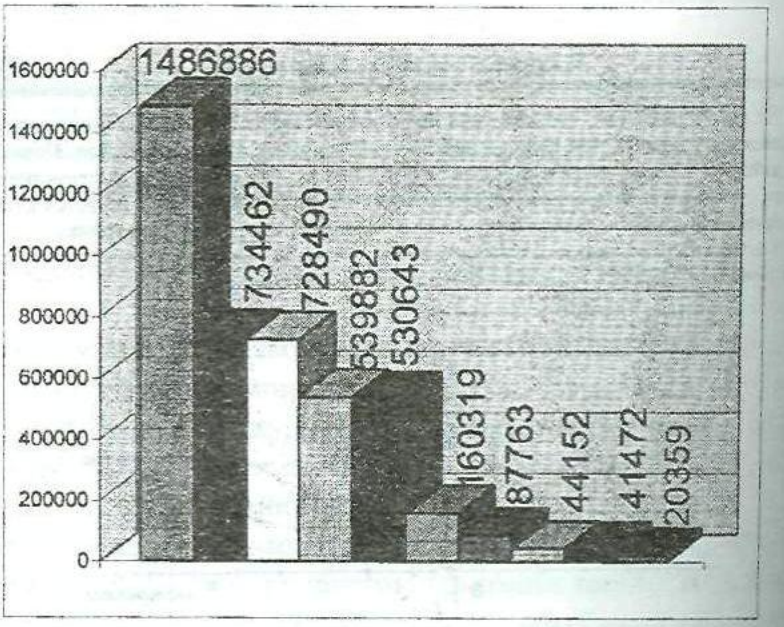

Рис. 2 Діаграма розмірів відсканованої сторінки

Огляд формату DjVu

Новий метод стиску графічних зображень, названий DjVu, створений спеціально для держання високоякісних копій відсканованих кольорових документів з високим ступенем стиску. Що дозволяє потім швидку передачу образа документа через мережу Інтернет в умовах ліній з низькою швидкістю передачі даних (що особливо актуально для телекомунікаційних каналів країн СНД) відтворювати візуальну копію документа включаючи колір, шрифт, картинки і текстуру папера. Метод DjVu розподіляє 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 стиску: дворіневий передній план зображення кодується за схемою AT&T у новий JBIG2 факс стандарт, а для фону і малюнків використовується новий метод хвильового стиску IW44. Обидва методи використовують новий адаптивний арифметичний кодер, називаний Z-кодером. Звичайна кольорова сторінка журналу формату А-4 із роздільною здатністю 300 dpi (dots per inch) може бути стиснута до розмірів 40-60Кб, що приблизно в 5-10 разів менше, ніж зображення у форматі JPEG при однакових рівнях ступеня стиску (див. рис. 2). Версія декодера зображення DjVu реалізована у вигляді програмного додатка для всіх популярних веб браузерів Internet Explorer, Netscape, Opera [1, 2].

Стиск повнокольорової інформації про документ формату А4 до розміру середньої Web-сторінки (46 Кбайт за даними на 1999 р.) 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 роздруківку 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 кругла пляма від винної пляшки поруч с 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 достатній стиск, однак вузько спеціалізовані або на фотографіях, або на чорно-білій графіці й тексті. Тому змішані зображення у їх виконанні виглядають таким, що важко прочитати. Розробники DjVu врахували негативний досвід створення "універсального солдата", їх розробка 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 розрізненням не меншим за 300 dpi.

Навпаки, невеличке розмиття фонової графіки навіть поліпшує сприйняття тексту. Тому фон без втрать для загального враження зберігається з розрізненням 100 dpi в окремому шарі ("background").

Основна проблема – відокремити текст від фону, особливо якщо це кольоровий текст, й крім того, різнокольоровий. Здебільшого колір тексту в документах практично однаковий у межах одного знаку. Це дозволяє зберігати кольорову інформацію про текст с розрізненням всього 25 dpi (шар "foreground") див. табл. 1).

Таблиця 1. Розподіл у файлі формату DjVu

Шар	Пояснення	Розрізненість, dpi	Глибина кольору, bits/pix
Mask	Монохромна маска–трафарет	300	1
Background	Кольорове тло	100	24
Foreground	Кольори маски	25	24

Розподіл зображення на текст та тло (формування шару-маски) базується на так званій мультимасштабній кластеризації. Зображення розбивається на рівномірні вкладені сітки, в кожній комірці котрих відбувається розпізнавання текстових та фонових кольорів за максимальними пікам на гістограмі. Відокремивши текст від фону у найкрупнішій сітці, алгоритм переходить до уточнення на базі даних з сіток меншого розміру.

В DjVu для стиску фона, 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 в 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 багато застосувань: онлайнові книжкові магазини, картографічна інформація й навіть е-хіромантія, де надіслана за поштою фотографія долоні обробляється подібним чином.

Наразі цей формат вже здобув широке застосування у бібліотеках різного профілю, в релігійних організаціях, у радіоаматорських колах, використовуючи формат DjVu, видаються онлайнові 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, й 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 та грошей.

Досить об'єктивна оцінка якості в порівнянні з вже відомими форматами показала, що незначне погіршення якості на кольорових зображеннях повністю компенсується ступенем стиску, котрий досягає сотень й тисяч разів, а на чорно-білих зображеннях 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 свій вибір ми зупинили на форматі DjVu, котрий розробляється американською компанією LizardTech. Формат DjVu реально дозволяє здійснити надто велике стискання зображень високої роздільності.

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 Kb для чорно-білого зображення формату А4 з роздільністю 300 dpi; 80–100 Kb для такого ж повнокольорового зображення);

повне збереження виду видання;

орієнтація на середовище Internet/Intranet й простота забезпечення навігації всередині публікації [2].

Недоліком такого рішення є те, що сторінка у DjVu–форматі є зображенням й 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 Internet існують безкоштовні пакети, котрі дозволяють відтворювати повноцінні зображення у форматі DjVu, включаючи й багатосторінкові файли, перелік таких пакетів наведено в табл. 2.

Таблиця 2 – Програмні продукти для створення файлів формату DjVu

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	Утиліта командного рядка для перетворення односторінкових документів формату DjVu
DjVuWebBrowser	1805	Перегляд документів формату DjVu безпосередньо в броузерах.

Створення файлів у форматі DjVu

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

Алгоритм технології створення файлу формату DjVu наведено на рис. 3.

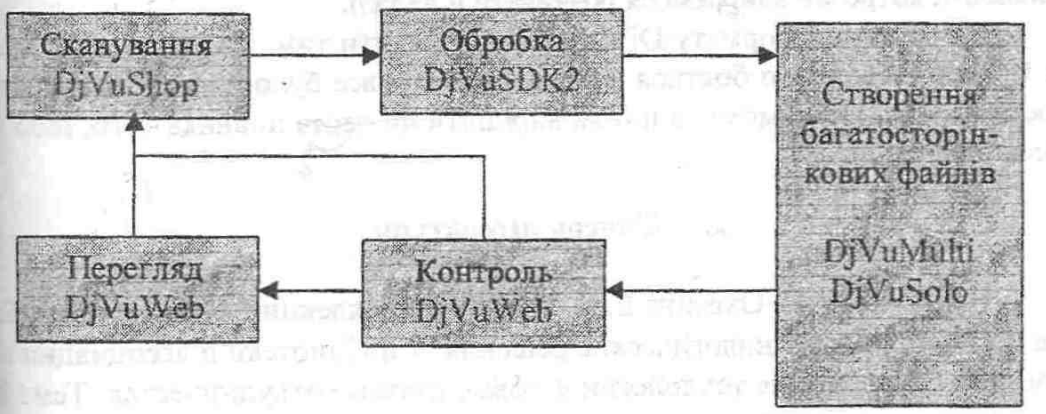

Рис. 3. Технологія створення файлу формату DjVu

Сканування твердої копії та створення власне DjVu файлу здійснюється за допомогою програми DjVuShop 2.0, котра зарекомендувала себе з кращого боку ніж програма DjVuSolo 3.1 за часом створення файлу формату DjVu з використанням комп'ютерів Pentium 120 – 150.

За необхідності використовується обробка відсканованих файлів обертанням їх на необхідний кут. Ця операція здійснюється за допомогою утиліти командного рядка DjVuSDK2.

Після сканування та обробки всього необхідного матеріалу здійснюється створення багатосторінкових файлів. Це може бути як один багатосторінковий файл, котрий містить всі відскановані файли (операція bundle), так і створений додатково індексний файл (операція join). Обидві ці операцій можна виконати, використовуючи або утиліти командного рядка DjVuBundle та DjVuJoin з пакету DjVuMulti 3.0, або програма DjVuSolo 3.1, котра дозволяє здійснювати ці операцій у віконному режимі.

Через те, що створення багатосторінкового файлу є дуже громіздка робота, то необхідним є візуальний контроль створеного файлу, котрий здійснюється або програмою DjVuSolo 3.1, переглядом за допомогою будьякого броузера, до котрого додано програму DjVuWebBrowser.

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

Висновки

Виходячи з усього вище викладеного в бібліотеці Харківського національного університету створюються повнотекстові електронні документи на базі формату djvu.

Але залишається ще одна проблема: доставка електронних документів читачеві.

Випробувані технологічні процеси по створенню файлів формату DjVu довели свою простоту, ефективність та економічність.

Для нас створення цифрових колекцій стало одним з важливих напрямків діяльності, котре ми збираємося розвивати й надалі.

Використання формату DjVu можна порадити тим, хто цінує час, але жадає й якості. Не варто боятися нового формату, все було колись новим, краще просто спробувати самому, а потом вирішити чи варта новинка того, щоб її застосовувати.

Список літератури

1. Негуляев Е. А., Охезина Е. А. Цифровые коллекции в вузовской библиотеке. Концепция и технологические решения. // 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 медунар. конф. "Крым 2002". – М.: ГПНТБ России, 2002. Т. 1 – С. 271 – 274

2. Волохін О. "Дежа вю" – нові технології для цифрової бібліотеки // Бібліотечна планета, 3, 2001, С. 9 – 11

Аннотация

Создание полнотекстовых документов в формате djvu

В статье приведены сравнения графических форматов по объему, представлено описание нового графического формата DjVu. Основываясь на опыте создания полнотекстовых электронных документов в формате DjVu приведена технология изготовления таких документов с использованием соответствующих программных пакетов.

Abstract

Development of full-text document in DJVU format

The article treats the issues of comparison of graphic formats by their volumes describer a new graphic format, the DjVu. The experience gained from designfull text electronic documents in the DjVu format suggests the techniques that be applied to design such documents
using the respective software.