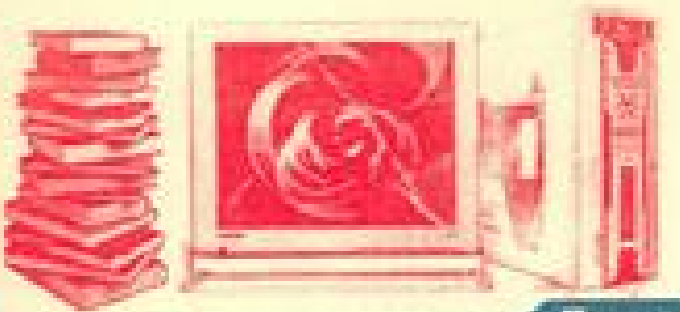
Полтавський університет
споживчої кооперації України

ІНФОРМАЦІЙНІ РЕСУРСИ та ПОСЛУГИ: нові види, проблеми розвитку та використання

Матеріали міжнародної науково-практичної конференції

8-9 червня 2004 року



Полтава
РВВ ПУСКУ

руси и Республиканская научно-техническая библиотека смогли бы управлять использованием такой информации. РНТБ – патентной информацией, а ЦНБ НАН Беларуси научной литературой. При этом обязательным условием такой работы следует также считать организацию обучения потребителей информации, то есть открыть специализированный учебный центр.

СТВОРЕННЯ ПОВНОТЕКСТОВИХ ДОКУМЕНТІВ ДЛЯ ЕЛЕКТРОННОЇ БІБЛІОТЕКИ

Влащенко Л.Г.;
Грищенко Т.Б.;
Нікітенко О.М., к.т.н.
Харківський національний університет радіоелектроніки

Одним з найважливіших показників зміни способу 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 і в системі бібліотечного сервісу. Зникла необхідність у багатьох інформаційних формах бібліографічної діяльності й розвинулися зовсім нові форми пошуку та зберігання учбової 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 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

ції й бібліотеки вже оцифровували більшість рукописів, які мають
-небудь цінність, однак розміри отриманих файлів не дозволяють
йомитися з ними через Інтернет.
3 грудня 1997 року у науковій бібліотеці Харківського національ-
го університету радіоелектроніки (ХНУРЕ) було розпочато роботи
творення своїх власних цифрових ресурсів. При цьому головними
огами, що висувалися, були простота, зручність та ефективність
нологічного процесу виготовлення електронних копій.
Ми бачимо, що вищезгадана 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 склається 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 документа). Час сканування залишається сталим і залежить тільки від типу використовуваного сканера. 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 найменше часу 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 більше часу, ніж попередні. Найдовшою є обробка відсканованого матеріалу у текстових форматах, через те, що після сканування і розпізнавання необхідно виправити помилки сканування.
Відомо, що зберігання відсканованого матеріалу потребує певних обсягів запам'ятовуючих пристроїв, тому при створенні повнотекстових документів треба враховувати й цей чинник. Для з'ясування розмірів форматів було відскановано одну сторінку формату А4 з науково-технічного журналу.
При цьому умови сканування були такі: формат А4, розрізнювальна здатність *300 pdi*, режим сканування *line art*.

На рисунку 1 наведено діаграму розмірів відсканованої сторінки. З цього рисунка випливає, що за розмірами найбільш придатним форматом є *djvu*.

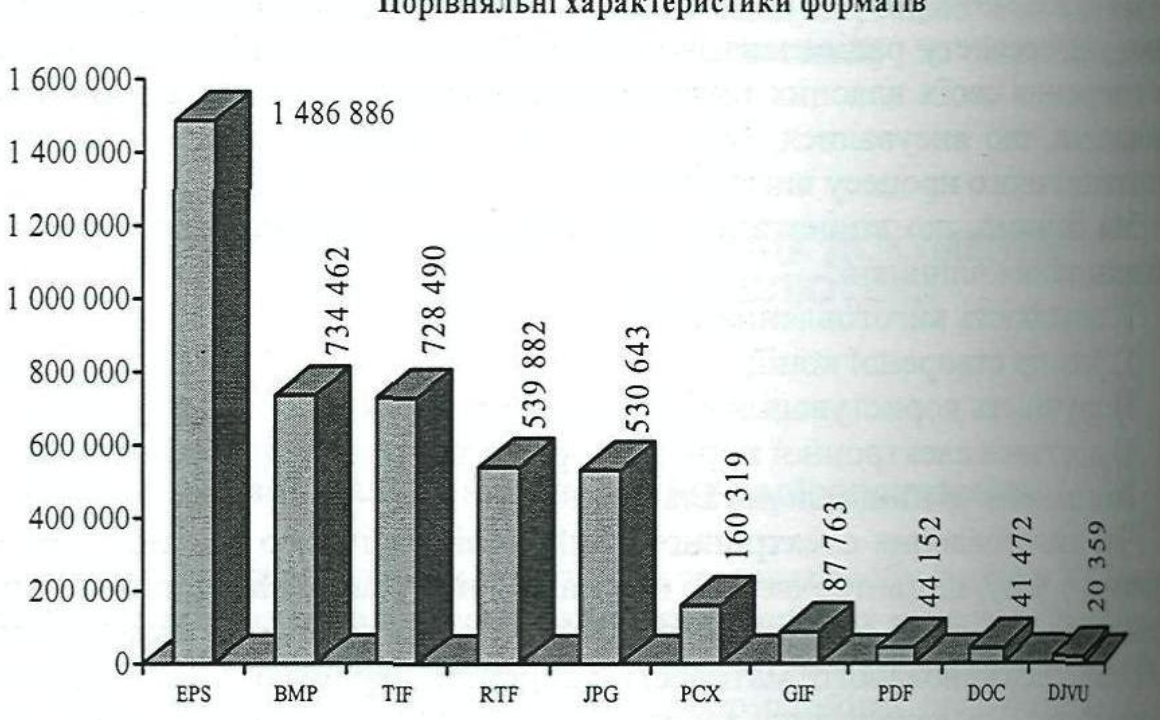


Рис. 1. Діаграма розмірів відсканованої сторінки

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

Наразі здійснюються активні розробки нового графічного формату, який би задовольняв потреби мережі. Однією з перших закінчених розробок є формат *DjVu* фірми *AT@T*, який 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*, що є безкоштовними для некомерційного використання. Засобом перегляду таких файлів може бути звичайний *web*-браузер. Було здійснено роботи з освоєння технології створення документів формату *DjVu*.

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 умо-

ліній з низькою швидкістю передачі даних (що особливо актуаль-для телекомунікаційних каналів країн СНД) відтворювати візуаль-копію документа включаючи колір, шрифт, картинки і текстуру еру. Метод *DjVu* розподіляє зображення на текст і малюнок. Для ту необхідний більш високий ступінь роздільної здатності, на міну від малюнків та фону зображення. Малюнок же звичайно ьш однорідний за своїм складом, тому може бути закодований з шим ступенем роздільної здатності. Потім застосовується новий од для максимізації ступеня стиску: дворівневий передній план раження кодується за схемою *AT&T* у новий *JBIG2* факс-стандарт, ля фону і малюнків використовується новий метод хвильового ску *IW44*. Обидва методи використовують новий адаптивний ариф-ичний кодер, названий *Z*-кодером. Звичайна кольорова сторінка рналу формату А4 із роздільною здатністю *300 dpi* (*dots per inch*) же бути стиснута до розмірів 40-60 Кб, що приблизно в 5-10 разів ще, ніж зображення у форматі *JPEG* при однакових рівнях ступеня ску (рис. 1). Версія декодера зображення *DjVu* реалізована у виг-програмного додатка для всіх популярних *web*-браузерів *Internet lorer*, *Netscape*, *Opera*.

Стиск повнокольорової інформації про документ формату А4 до міру середньої *web*-сторінки (46 Кб за даними на 1999 р.) теоретич-можливий. Враховуючи зростаючу суспільну потребу в доступі до гіналів», здається дивним, що стандарт на графічні зображення ого призначення формується тільки сьогодні.

Формат *DjVu* – перший крок до «кольорового факсу» і його кентовано на передачу, перегляд у мережі й роздруківку переважно стових документів, для яких важливе значення має не тільки зміст, й форма: колір та фактура пергаменту, відірваний куточок і сліди складання вчетверо, пляма після підпису й кругла пляма від винної шки поруч з печаткою. Архіви всього світу накопичили величезну кість історичних паперів з неповторним колоритом такого чину.

Існуючі компактні формати *JPG*, *GIF*, факс-стандарт *CCITT* та G забезпечують достатній стиск, однак вузько спеціалізовані або фотографіях, або на чорно-білій графіці й тексті. Тому змішані раження у їх виконанні виглядають таким, що важко прочитати. обники *DjVu* врахували негативний досвід створення «універсаль-солдата», їх розробка складається з трьох форматів «в одному рі». Розділ «обов'язків» всередині *DjVu* базується на простих ереженнях та фактах.

кст та інші контрастні малюнки зручно читати при скануванні з нюванням не меншим за *300 dpi*.

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

Основна проблема – відокремити текст від фону, особливо якщо це кольоровий текст, і крім того, різнокольоровий. Здебільшого колір тексту в документах практично однаковий у межах одного знака. Це дозволяє зберігати кольорову інформацію про текст с розрізненням всього *25 dpi* (шар «*foreground*») (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 різнорозмірні 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 із сіток меншого розміру.

В *DjVu* для стиску фону, маски та кольорової інформації про маску застосовують різні алгоритми. Фон стискається вейвлет-алгоритмом *IW44* (4 × 4 *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 багато застосувань: онлайнові книжкові магазини, картографічна інформація й навіть *e*-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
яг у форматі *DjVu* біля 30 Кб, а у кольорі біля 60 Кб, то стає
зумілою економія часу та грошей.

Досить об'єктивна оцінка якості в порівнянні з уже відомими
рматами показала, що незначне погіршення якості на кольорових
раженнях повністю компенсується ступенем стиску, котрий сягає
ень й тисяч разів, а на чорно-білих зображеннях практично не видно.
жливі конкуренти у вигляді *tiff, gif, jpg* сильно програють в обсязі.

До речі, популярний *jpg* зовсім непридатний для чорно-білих
сканованих зображень.

Саме на чорно-білих схемах і текстах, а поліграфічні видання
більшого такими і є, перевага *djvu* є колосальною.

Таким чином свій вибір ми зупинили на форматі *DjVu*, який
робляється американською компанією *LizardTech*. Формат *DjVu*
ально дозволяє здійснити надто велике стискання зображень високої
здільності.

Основні переваги цього формату:

- доступ до цифрової колекції по мережі Інтернет/Інтернет з
користанням стандартного програмного забезпечення (необхідно
ше встановити додатковий модуль для браузера);
- висока якість та малий об'єм зображень будь-яких видів
30 Кб для чорно-білого зображення формату А4 з роздільніс-
300 dpi; 80-100 Кб для такого ж повнокольорового зображення);
- повне збереження виду видання;
- орієнтація на середовище Інтернет/Інтернет й простота забезпе-
ння навігації всередині публікації.

Недоліком такого рішення є те, що сторінка у *DjVu*-форматі є
браженням і не дозволяє використовувати будь-який пошук. Власне,
формат *DjVu* дозволяє зберігати текстову інформацію після
оцедури *OCR* (оптичного розпізнавання символів).

Зрозуміло, що бібліотеки мають обмаль коштів, яких ледве вистача
на комплектування, тому бажано мати програмні пакети за мініль-
ною ціною. На щастя, в мережі Інтернет існують безкоштовні па-
ти, які дозволяють відтворювати повноцінні зображення у форматі
Vu, включаючи й багатосторінкові файли, перелік таких пакетів
ведено в таблиці 2.

блиця 2. Програмні продукти для створення файлів формату *DjVu*

Назва	Обсяг, Кб	Призначення
jVuShop 2.0	1613	Створення, редагування та перегляд односторінкових документів формату *DjVu*
jVuSolo 3.1	2228	Створення, редагування та перегляд одно- й багатосторінкових документів формату *DjVu*

Назва	Обсяг, Кб	Призначення
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

Методика створення файлів у форматі DjVu

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

Алгоритм створення файлу формату *DjVu* наведено на рисунку 2.

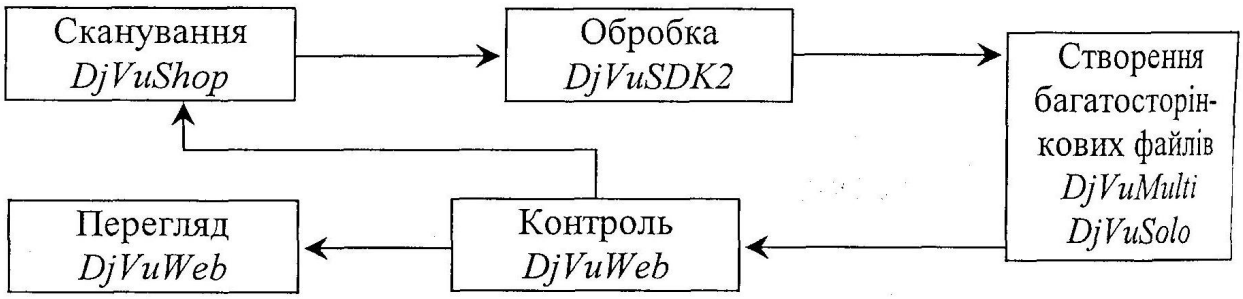


Рис. 2. Методика створення файлу формату DjVu

Сканування твердої копії та створення власне *DjVu*-файлу здійснюється за допомогою програми *DjVuShop 2.0*, яка 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 *DjVu*Multi 3.0, або програма *DjVuSolo 3.1*, яка дозволяє здійснювати ці операцій у віконному режимі.

Через те, що створення багатосторінкового файлу – громіздка робота, необхідним є візуальний контроль створеного файлу, який здійснюється або програмою *DjVuSolo 3.1*, переглядом за допомогою будь-якого броузера, до котрого додано програму *DjVuWebBrowser*.

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

Висновки

Виходячи з усього вищевикладеного в бібліотеці Харківського ціонального університету створюються повнотекстові електронні кументи на базі формату *djvu*.

Але залишається ще одна проблема: доставка електронних доку-нтів читачеві.

Випробувані технологічні процеси створення файлів формату *DjVu* вели свою простоту, ефективність та економічність.

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

Використання формату *DjVu* можна порадити тим, хто цінує час, прагне і якості. Не варто боятися нового формату, все було колись вим, краще просто спробувати самому, а потім вирішити, чи варта винка того, щоб її застосовувати.

БІБЛІОТЕКА ВИЩОГО НАВЧАЛЬНОГО ЗАКЛАДУ В СИСТЕМІ ДИСТАНЦІЙНОГО НАВЧАННЯ

ищенко Т.Б.;
ітенко О.М., к.т.н.
рківський національний університет радіоелектроніки

Постіндустріальне або інформаційне суспільство відрізняється вик-чно швидким розвитком інформаційних та комунікаційних техно-ій, можливості яких стають безпрецендентними для ефективного рішення багатьох професійних, економічних, соціальних та побуто-х проблем. Застосувати ці можливості можуть лише ті члени спільства, котрі матимуть необхідну компетентність, що дозволяє ентуватися в новому інформаційному просторі, зберігаючи свою обутність, використовувати переваги глобалізації. Мова йде про бхідність зміни змісту освіти на базі нових можливостей освітньо-середовища, про засвоєння інформаційної культури, яку розуміють вищий вияв людської освіченості, включаючи особистнісні якості професійну компетентність.

Інформаційно грамотна людина визначається, як така, котра здатна зуміти, коли їй потрібна інформація, яка вміє її знаходити, оцінюва-ти та ефективно використовувати.