ZESZYTY NAUKOWE POLITECHNIKI ŚLĄSKIEJ 2006
Seria: TRANSPORT z. 62 Nr kol. 1721

Mariusz RYCHLICKI[1]

EFEKTYWNA TECHNIKA ELEKTRONICZNEJ ARCHIWIZACJI DOKUMENTÓW PRZEDSIĘBIORSTWA TRANSPORTOWEGO

Streszczenie. W artykule zwrócono uwagę na rosnące znaczenie systemów elektronicznego przetwarzania dokumentów w funkcjonowaniu współczesnych przedsiębiorstw. Podkreślono rolę, jaką w tym kontekście odgrywają techniki kompresji. Sformułowano kryteria i oczekiwania, pozwalające na wybór spośród nich techniki najbardziej efektywnej. Zwrócono uwagę na standard DjVu, nie tylko spełniający te oczekiwania, ale oferujący wiele dodatkowych możliwości. Zaproponowano koncepcję archiwizacji dokumentów przedsiębiorstwa, w szczególności transportowego, w oparciu o istniejące narzędzia tworzenia dynamicznych witryn WWW. Wykazano, że rozwiązanie takie może być alternatywą dla specjalizowanego oprogramowania, szczególnie z uwagi na wyjątkowo niskie koszty wdrożenia. Przedstawiono także drugą koncepcję, zakładającą wykorzystanie profesjonalnego systemu obsługi bibliotek cyfrowych.

EFFECTIVE TECHNIQUE OF TRANSPORT COMPANY ELECTRONIC DOCUMENT ARCHIVING

Summary. In the paper attention was paid to the growing importance of electronic document processing systems in functioning of present-day enterprises. The role the compression techniques in this context plays was underlined. Criteria and expectations were formulated that enable selection of a most effective technique from among them. Attention was paid to the DjVu standard that not only fulfils these expectations but offers a series of additional possibilities. An archiving concept of enterprise documents archiving was proposed, in particular of transport one, on the basis of existing tools of dynamic WWW sites creation. It was proved that such a solution can be an alternative for dedicated software, in particular considering exceptionally low implementation costs. A second concept was also presented that assumes use of a digital libraries professional service system.

1. WPROWADZENIE

Techniki kompresji danych, polegające na zmniejszeniu objętości zbioru źródłowego, należą do najtańszych metod oszczędności miejsca pamięci i nośników systemów teleinformatycznych. Nie bez znaczenia jest także pośredni wpływ na efektywną szybkość samej transmisji danych – mniejsza ilość danych, to krótszy czas ich przesłania łączem o danej przepustowości. Tym samym zrozumiałe jest, że rozwój technik kompresji i systemów teleinformatycznych dawno stał się ściśle związany oraz wspólnie uwarunkowany.

Duże sukcesy w dziedzinie rozwoju technik kompresji, wspierane wzrostem mocy obliczeniowej systemów komputerowych oraz rozwojem technologii wytwarzania pamięci i

[1] Wydział Transportu, Politechnika Warszawska, Koszykowa 75, 00-662 Warszawa, mry@it.pw.edu.pl

nośników danych przy jednoczesnym radykalnym spadku ich ceny, przyczyniły się do rozwoju wielu nowych obszarów ich zastosowań. Jednym z nich jest upowszechnienie i urzeczywistnienie idei elektronicznej wymiany danych EDI (ang. *Electronic Data Interchange*), gdzie zdecydowana większość (docelowo wszystkie) dokumentów powstaje i jest przetwarzana jedynie w postaci elektronicznej.

Korzyści z takiego rozwiązania są oczywiste:
- szybkość wytworzenia,
- łatwość przechowywania,
- prostota wyszukiwania i przeszukiwania,
- efektywność przesyłania,
- niskie koszty archiwizacji.

Poza oferowanymi korzyściami, rozwojowi tej idei sprzyjają nowe uwarunkowania prawne, które stale poszerzają zakres stosowania dokumentów elektronicznych na równi z ich papierowymi odpowiednikami [1]. Dzięki temu wizja „biura bez papieru" stała się już dla wielu przedsiębiorstw rzeczywistością.

Niemniej stale pozostaje wiele firm, głownie małych i średnich, które nie są w stanie wprowadzić tej idei w życie we własnym przedsiębiorstwie. Głównym problemem jest tutaj bariera ekonomiczna, uniemożliwiająca dostosowanie systemu teleinformatycznego do wymagań EDI zarówno w zakresie sprzętu, jak i specjalizowanego oprogramowania systemowego. Wśród tych firm znajdują się także przedsiębiorstwa transportowe (kurierskie, przewozowe, spedycyjne), które w obliczu różnorodności oraz mnogości kontrahentów stają przed coraz większymi problemami archiwizacji i przetwarzaniu dokumentów w ich papierowej postaci. Wiele z tych problemów mogłoby zostać szybko i sprawnie rozwiązanych, gdyby tylko istniała spójna technika przetwarzania (głównie kompresji i archiwizacji) dokumentów, nie wymagająca znaczących nakładów finansowych.

2. EFEKTYWNA TECHNIKA PRZETWARZANIA DOKUMENTÓW

Poszukiwania odpowiedniego narzędzia, gwarantującego spełnienie stawianych wymagań, należy rozpocząć od ich ścisłego zdefiniowania. Warunek minimalizacji kosztów rozwiązania zawęża poszukiwania do rozwiązań opartych na popularnych systemach operacyjnych (Windows, Linux) i pozwalających w prosty sposób wykorzystać powszechnie dostępne platformy prezentacji dokumentów (np. HTML i pokrewne). Tym samym do głównych wymagań należeć będą:
- Praca w popularnych systemie operacyjnym (Windows, Linux).
- Wykorzystanie powszechnych (głównie internetowych) platform i narzędzi prezentacji dokumentów.
- Prostota przetwarzania dokumentu do formatu docelowego, zarówno z postaci papierowej, jak i elektronicznej.
- Wysoka wydajność – duże współczynniki kompresji, przy jednoczesnej wysokiej zgodności z oryginałem.
- Prostota tworzenia dokumentów zbiorczych.
- Łatwość przeglądania i przeszukiwania dokumentów archiwalnych.
- Wbudowany mechanizm rozpoznawania tekstu dla dokumentów skanowanych.
- Bezpośredni dostęp do dokumentów (stron) w dokumentach zbiorczych.

Przez szereg lat większości z tych wymagań był w stanie sprostać, wprowadzony przez firmę Adobe, format PDF (ang. *Portable Document Format*). Pojawiający się sporadycznie na rynku konkurenci nie byli w stanie zaoferować zdecydowanie lepszych ani nowych możliwości, co w połączeniu ze sprawną polityką firmy Adobe (np. bezpłatna przeglądarka) prowadziło jedynie do dalszego umocnienia pozycji formatu PDF. Należy

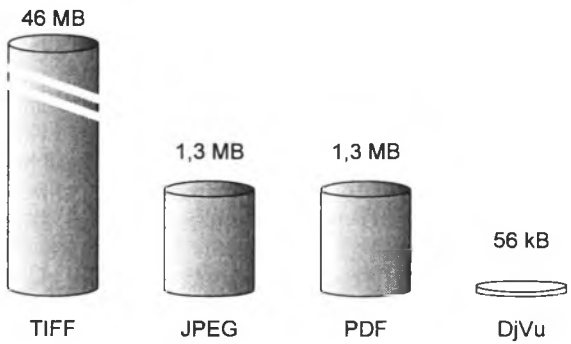
jednak pamiętać, że format PDF powstał z myślą o zastosowaniach do wydruków dokumentów, co znacząco ogranicza komfort ich przeglądania, np. bezpośrednio na ekranie monitora. Dodatkowym problemem jest brak możliwości zewnętrznego (spoza pliku) bezpośredniego odwołania do konkretnej strony dokumentu, co dodatkowo ogranicza zastosowania tego formatu w archiwizacji dokumentów i rozbudowanych (sieciowych) aplikacjach baz danych.

W ostatnim jednak czasie na rynku elektronicznego przetwarzania dokumentów pojawiło się narzędzie, którego twórcy wprost zapowiadają koniec ery formatu PDF. Jest nim standard DjVu, powstały wręcz z myślą o zastosowaniach w archiwizacji dokumentów i nie tylko pozbawiony dwóch podstawowych wad formatu PDF, lecz dodatkowo oferujący nowe możliwości przy lepszych parametrach kompresji.

3. STANDARD DJVU

Od roku 1996 w laboratoriach firmy AT&T trwają prace nad nową, wysoko efektywną techniką przetwarzania dokumentów elektronicznych. Prace te doprowadziły do powstania nowego formatu dokumentów, tzw. DjVu. Pod koniec lat 90. XX w., na podstawie umów licencyjnych i praw patentowych, amerykańska firma LizardTech Inc. wprowadziła ten format na rynek, implementując jego obsługę w stworzonym oprogramowaniu – rodzinie *Document Express*. W bardzo krótkim czasie okazało się, że możliwości tego formatu stanowią realne zagrożenie dla formatów Postscript i PDF, pretendując go do miana nowego standardu elektronicznego przetwarzania dokumentów. Proces ten trwa już kilka lat i choć początkowo niewidoczny, staje się coraz bardziej dynamiczny i zauważalny.

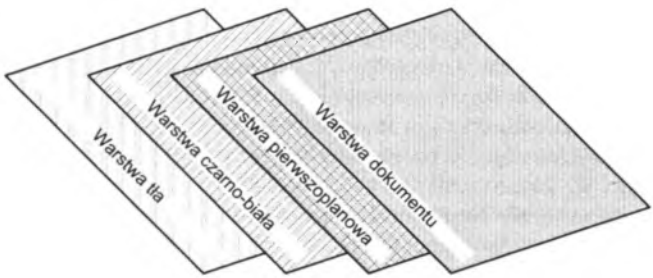
Przyczyn coraz większego sukcesu standardu DjVu jest wiele, jednak najważniejszą z nich jest bez wątpienia wysoka skuteczność i uzyskiwane duże współczynniki kompresji. Dokumenty zapisane w tym formacie są do 1000 razy mniejsze od plików TIFF i zwykle od 5 do 100 razy mniejsze od plików JPEG i PDF [2]. Porównanie wielkości zbiorów danych w tych formatach, powstałych po zeskanowaniu przykładowej kolorowej strony, zostało przedstawione na rysunku 1.



Rys. 1. Porównanie wielkości zbiorów danych [2]
Fig. 1. Comparison of data set sizes [2]

Tak wysoka skuteczność kompresji formatu DjVu została osiągnięta dzięki nowatorskiej metodzie segmentacji obrazu [3]. Jej podstawowa idea polega na rozdzieleniu przetwarzanych obrazów (bez względu na stopień ich złożoności) na warstwy (rysunek 2). Ich główny podział, to podział na warstwę pierwszoplanową i warstwę tła. Następnie warstwy te poddawane są odrębnej optymalizacji i kompresji, stosownie do reprezentowanej przez nich

treści i właściwości. Warstwa tła zawiera, zgodnie z nazwą, reprezentowane w niskiej rozdzielczości (zwykle 100dpi) tło dokumentu. Kodowane jest ono techniką typową dla obrazów płynnych przejść tonalnych (ang. *Continuous Tone Technique*). Warstwa pierwszoplanowa zawiera wysokiej (wyższej) rozdzielczości (zwykle 300dpi) maskę obrazu, definiującą szczegóły i kształty detali obrazu. Informacja o kolorze jest przetwarzana do postaci niskiej rozdzielczości (25dpi) obrazu pierwszoplanowego, którego kolory są wykorzystywane jako szablon dla maski obrazu. Niezależnie tworzone są czarno-białe warstwy tekstu, zarówno (w zależności od rodzaju dokumentu) jako wyodrębniony zestaw szczegółów obrazu, jak i zestaw znaków odczytanych techniką rozpoznawania tekstu OCR (ang. *Optical Character Recognition*). Ta ostatnia właściwość techniki DjVu ma fundamentalne znaczenie w dalszym przetwarzaniu, głównie przeszukiwaniu i katalogowaniu dokumentów DjVu.



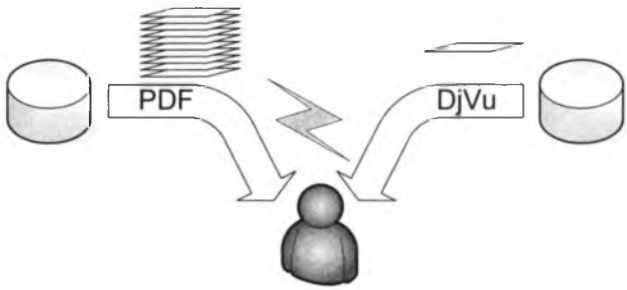
Rys. 2. Podstawowa idea podziału dokumentu na warstwy
Fig. 2. Basic idea of document division into layers

W rezultacie złożonego i wielowarstwowego przetwarzania dokumentu plik formatu DjVu jest nie tylko znacznie mniejszych rozmiarów, niż jego odpowiedniki w innym standardzie, np. PDF. Ma on wiele dodatkowych cech, które czynią go niezwykle interesującym dla systemów elektronicznego przetwarzania dokumentów:

- **Jakość**: wysoki współczynnik kompresji nie jest ani kompromisem ani kosztem w odniesieniu do jakości, lecz rezultatem zastosowanej metody.
- **Przenośność**: pliki formatu DjVu są plikami rastrowymi, dzięki czemu mogą być swobodnie przenoszone pomiędzy różnymi platformami systemowymi (Windows, Sun/Solaris, Linux, Unix) bez typowych dla innych rozwiązań problemów, związanych z różnorodnością stosowanych czcionek.
- **Efektywność**: przeglądarki plików DjVu przetwarzają w danym momencie jedynie aktywny (oglądany) fragment dokumentu, co pozwala na uzyskanie maksymalnej szybkości jego przetwarzania (powiększania, przesuwania).
- **Optymalność**: format DjVu został zoptymalizowany pod kątem zastosowań w sieciach rozległych, głównie dla zapewnienia niemal natychmiastowego odczytu danych, co osiągnięto poprzez zapewnienie dostępu do aktualnie interesującego fragmentu pliku, bez konieczności wcześniejszej jego transmisji w całości.
- **Podatność**: wysoka jakość dokumentów DjVu pozwala na poddanie ich przetwarzaniu przy wykorzystaniu innych technik, od rozpoznawania tekstu OCR, przez zbiorcze indeksowanie i wyszukiwanie, aż do eksportu do innych formatów, np. TXT lub XML.

Jak widać z powyższego zestawienia, standard DjVu jest w stanie spełnić wszystkie wymagania, stawiane (i sformułowane w p. 2) efektywnej technice przetwarzania dokumentów. Na szczególne podkreślenie zasługuje cecha formatu, gwarantująca jego

optymalne wykorzystanie w sieciach rozległych. To dzięki niej pliki formatu DjVu mogą być bezpośrednio osadzane w dokumentach HTML zamiast plików obrazu lub jako ich odrębne uzupełnienie. Najważniejszą konsekwencją tej cechy jest fakt, że użytkownik uzyskuje natychmiastowy dostęp do interesującego go fragmentu dokumentu bez konieczności transmisji całego pliku, jak ma to miejsce przy zastosowaniach klasycznych (rysunek 3). Dzięki temu wielkość pliku przestała mieć wpływ na szybkość przeglądania dokumentu, co znosi ograniczenia ich rozmiaru w kontekście udostępniania ich w sieci.



Rys. 3. Różnica w dostępie do strony dokumentu PDF i DjVu
Fig. 3. Difference in access to the PDF and DjVu document page

Samo prezentowanie dokumentu nie ogranicza się bynajmniej do jego udostępnienia. Standard oferuje wiele opcji, które pozwalają dynamicznie wpływać na sposób wyświetlania dokumentu za pomocą zestawu argumentów na sposób zbliżony do CGI-Bin. Sam użytkownik ma także możliwość skonfigurowania przeglądarki w taki sposób, by dokumenty prezentowane były w ulubionej przez niego postaci [4]. Na komfort pracy z dokumentami DjVu wpływają także m.in. rozbudowane możliwości tworzenia miniatur stron, spisów treści oraz osadzania hiperłączy, prowadzących do innych stron, dokumentów lub plików.

4. MOŻLIWOŚCI STWORZENIA ARCHIWUM DOKUMENTÓW

Cechy dokumentów formatu DjVu, co zaprezentowano wcześniej, pozwalają na wyjątkową łatwość ich integracji z rozwiązaniami sieciowymi, głownie internetowymi. Umożliwia to proste i szybkie stworzenie archiwum dokumentów w oparciu o istniejące narzędzia programistyczne. Co istotne, narzędzia te są nie tylko dostępne w bardzo szerokim wyborze, lecz także przy bardzo niskim koszcie zakupu licencji, a często wręcz bezpłatnie. Za ten stan rzeczy odpowiada zmierzch ery statycznego HTML i niezwykła, stale rosnąca, popularność dynamicznych witryn WWW [5]. Mechanizmy bazodanowe i techniki skryptowe działające po stronie serwera przestały być domeną portali i sklepów internetowych, a stały się nieodzownym składnikiem każdej, rozbudowanej strony WWW. Wśród narzędzi programistycznych od dawna wiedzie tu prym zestaw PHP i MySQL, pracujący pod kontrolą Apache. Dzięki temu stworzenie archiwum przedsiębiorstwa można byłoby powierzyć osobie odpowiedzialnej za obsługę serwisu internetowego firmy. Oczywiście, od rodzaju tworzonego archiwum i natury przechowywanych dokumentów zależeć będzie, czy będą one udostępniane w sieci jedynie wewnątrz przedsiębiorstwa, czy też publicznie. Dostępne techniki autoryzacji i identyfikacji nie wykluczają powstania rozwiązania mieszanego, opartego na hierarchicznej strukturze dostępu do dokumentów.

Oczywiście, w pełni funkcjonalne archiwum dokumentów musi oferować funkcję ich przeszukiwania pod kątem potrzeby znalezienia określonej zawartości. Procedura indeksacji dokumentów jest zazwyczaj procesem bardzo czasochłonnym. Jednak i tutaj przychodzi z pomocą standard DjVu, dla którego oferowane jest bezpłatne narzędzie przeszukiwania kolekcji dokumentów DjVu po ich zawartości warstwy tekstowej. Narzędziem tym jest IFilter, który integruje się z systemem operacyjnym, umożliwiając wyszukiwanie wprost z poziomu systemu.

Analizując jednak funkcje, jakie musi spełniać elektroniczne archiwum dokumentów, trudno nie oprzeć się wrażeniu, że funkcjonalnie w niczym nie odbiega ono od tzw. cyfrowej biblioteki. Musi przecież:

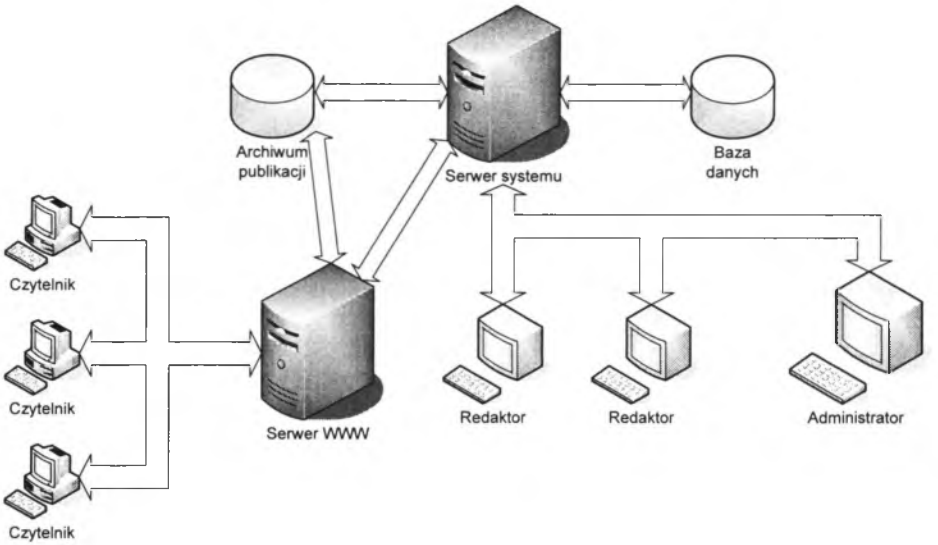
- Zostać stworzone i nadzorowane przez kompetentną osobę – Administratora.
- Gromadzić dokumenty w bazie danych i tworzyć archiwum publikacji.
- Przyjmować dodatkowe dokumenty i pozwalać na modyfikację istniejących uprawnionym użytkownikom (pracownikom przedsiębiorstwa) – Redaktor.
- Udostępniać dokumenty w sieci przedsiębiorstwa i/lub Internet.
- Pozwalać na dostęp uprawnionych użytkowników (pracowników) poprzez sieć – Czytelnicy.

Pierwszym, nadal rozwijanym, polskim projektem biblioteki cyfrowej jest środowisko dLibra. Pierwsze badania, prowadzące do jego powstania, realizowane były przez PCSS (Poznańskie Centrum Superkomputerowo-Sieciowe) już w 1996 roku. Obecnie system dLibra jest najpopularniejszym tego typu oprogramowaniem w Polsce, którego grono użytkowników stale się poszerza. Są to głównie biblioteki akademickie i publiczne, za czym przemawia fakt niskiego kosztu wdrożenia dla placówek państwowych. Nic nie stoi jednak na przeszkodzie, by z projektu tego nie mogły skorzystać przedsiębiorstwa czysto komercyjne.

Oparte na dLibrze systemy oferują użytkownikom nie tylko wszystkie przedstawione wyżej funkcje, ale oferują znacznie szersze możliwości [6]. Integrują w jednym środowisku nie tylko dokumenty wszystkich głównych formatów (HTML, DOC, PDF, DjVu itd.), ale także pliki audio i wideo [7]. Sama wymiana danych odbywa się w oparciu o powszechnie znane i docenione protokoły oraz standardy, do których należą RSS, RDF, MARC, DublinCore czy OAI-PMH.

Sam system jest systemem wielowarstwowym (rysunek 4), w którym dają się wyróżnić, typowe dla sieciowych rozwiązań bazodanowych, główne moduły: Serwera i jego Administratora oraz Redaktora i Czytelnika. Ich funkcje dokładnie odpowiadają sprecyzowanym wcześniej funkcjom elektronicznego archiwum dokumentów. Wymagania dotyczące sprzętu nie są wysokie, a samo oprogramowanie potrzebne do jego uruchomienia (oprócz głównej licencji dLibra) jest wręcz darmowe [6]:

- System operacyjny - Linux,
- Baza danych – PostgreSQL,
- Serwer aplikacji - Apache Tomcat,
- Wirtualna maszyna Javy - JRE Sun Microsystems Inc.



Rys. 4. Architektura systemu dLibra [6]
Fig. 4. dLibra system architecture [6]

Zatem uzyskujemy dwa sposoby stworzenia elektronicznego archiwum dokumentów przedsiębiorstwa przy minimalnych nakładach finansowych. W obydwu przypadkach wymagany będzie zakup oprogramowania do konwersji dokumentów DjVu. Jednak jego koszt w wersji podstawowej *Document Express Professional* jest na tyle niski, że nie stanowi praktycznie żadnej bariery finansowej. Drugie rozwiązanie wymagać będzie dodatkowo zakupu licencji środowiska dLibra. Nie jest to jeszcze produkt czysto komercyjny, dlatego nie istnieje jego konkretny cennik. Jednak ze wstępnych ustaleń i zapewnień PCSS jego koszt powinien być zbliżony do ceny dobrej klasy i mocy komputera PC, co także nie tworzy żadnej bariery finansowej.

5. PODSUMOWANIE

Dynamiczny rozwój systemów teleinformatycznych sprzyja idei w pełni elektronicznego przetwarzania dokumentów. Jednak jej praktyczna realizacja natrafia jeszcze na wiele przeszkód. Obok nadal istniejących ograniczeń prawnych, główną barierą są wysokie koszty wdrożenia specjalizowanego i zintegrowanego oprogramowania systemowego. Dlatego uzasadnione jest poszukiwanie prostych i tanich rozwiązań, które mogłyby pomóc małym i średnim firmom w urzeczywistnieniu tej idei. Takim narzędziem bez wątpienia jest standard efektywnej kompresji DjVu. Jego optymalizacja pod kątem wykorzystania w sieciach rozległych sprzyja zastosowaniu szeregu bezpłatnych narzędzi programistycznych. Dodatkowo, otwiera nowe możliwości tworzenia elektronicznego archiwum dokumentów na bazie rozwiązań dedykowanych cyfrowym bibliotekom. Z tego względu bliższe zapoznanie się z tym standardem wydaje się uzasadnione w kontekście zastosowań w przedsiębiorstwach, nie tylko o profilu transportowym.

Literatura

1. Ustawa z dnia 18 września 2001 r. o podpisie elektronicznym.
2. Bottou L., Haffner P., LeCun Y.: Efficient Conversion of Digital Document to Multilayer Raster Formats, AT&T Labs – Research.
3. Haffner P., Bottou L., Lecun Y., Vincent L.: A General segmentation scheme for DjVu document compression, ISMM 2002, Sydney, Australia.
4. DjVu Browser Plug-in v. 6.0.0, LizardTech Inc., 2005.
5. Meloni J. C.: PHP, MySQL, Apache dla każdego. Helion, Gliwice 2005.
6. Materiały informacyjne środowiska dLibra, http://dlibra.psnc.pl
7. Parkoła T.: Podręcznik użytkownika środowiska dLibra. PCSS 2005.