DjVu i *DjVuLibre*

Jakub Wilk

Wydział Matematyki, Informatyki i Mechaniki Uniwersytetu Warszawskiego

15 listopada 2007 r.

- $>$ 90% informacji znajduje się na papierze.[1] Większość z nich nie jest dostępna w Internecie.
- Udostępnianie skanowanych dokumentów w sieci w konwencjonalnych formatach jest problematyczne:
  - czytelność $\approx$ wysoka rozdzielczość; formaty PNG, JPEG, PDF:
    - mają niski współczynnik kompresji,
    - ich dekodowanie jest pamięciożerne;
  - mamy do wyboru:
    - niewygodną nawigację: HTML + plik graficzny dla każdej strony skanu
    - plik PDF olbrzymich rozmiarów.
- Rozwiązanie: **DjVu** /deʒa vy/ — metoda kompresji obrazów i format dokumentu przeznaczone zwłaszcza do dygitalizacji dokumentów papierowych.

[1] ⟨http://djvuzone.org/wid/⟩

Zastosowanie:

- książki,
- czasopisma,
- gazety,
- rękopisy,
- dokumenty historyczne;
- głównie skany,
- także dokumenty elektroniczne.

Zalety:

- mocna kompresja,
- wygoda przeglądania pogodzona z niewielkimi rozmiarami plików,
- *lekkie* wtyczki do przeglądarek WWW — dostępne za darmo,
- format pliku o otwartej specyfikacji.

Przykład

Kazimierz Kuratowski, Andrzej Mostowski *Teoria mnogości* — dostępna w Bibliotece Wirtualnej Matematyki ICM-u:

- PDF:
  - ⟨`http://matwbn.icm.edu.pl/kstresc.php?wyd=10&tom=27`⟩,
  - 6 plików PDF,
  - rozdzielczość 600 dpi,
  - 147 stron A4,
  - 87,3 MiB ($\approx$ 600 KiB/stronę);
- DjVu:
  - 20,3 MiB ($\approx$ 140 KiB/stronę) – ponad 4 razy mniejszy od oryginału.

DjVuBitonal (DjVuText, JB2):

- dla obrazów:
  - czarno-białych (zwłaszcza tekstu) lub
  - o małej liczbie kolorów (duże jednolite obszary);
- kompresja:
  - z użyciem słownika powtarzających się kształtów,
  - 2 – 10× mocniejsza niż CCITT GroupIV (TIFF, PDF),
  - 5 – 30 KiB / stronę w 300 dpi,
  - stratna lub bezstratna.

DjVuPhoto (IW44):

- dla obrazów o płynnych przejściach barw (zdjęcia);
- kompresja:
  - falkowa,
  - niektóre piksele mogą być oznaczone jako nieistotne,
  - ≈ 2 razy mocniejsza niż JPEG;
- dekompresja:
  - mały narzut pamięci,
  - $\approx 3\times$ szybsza niż JPEG-2000,
  - możliwa postępowa wizualizacja,
  - możliwa wizualizacja obszaru bez dekompresji całego obrazu.

Przykład

JPEG
2081 bajtów

C44
2026 bajtów

DjVuLayered (DjVu, DjVuDocument):

- dla obrazów:
  - skanowanych w kolorze lub skali szarości,
  - zawierających oprócz tekstu — grafikę,
  - o niejednolitym tle;
- 2 warstwy:
  - tło — IW44 lub JPEG,
  - pierwszy plan — IW44 lub JPEG + maska JB2 lub MMR;
- kawałki IW44 mają zazwyczaj obniżoną rozdzielczość.

Przykład

www.gazetaprawna.pl PORADNIA RACHUNKOWA PONIE

Masz pytanie, przyślij e-mail – odpowiemy Ci w dodatku Księgowość i Podatki

ksiegowosc@infor.pl

Agnieszka Michalak
kierownik ds. księgowości w FPA Group

Piot
dyrektor oddziału regio zachód, członek zarządu PKF

Czy w rachunkowości wykazywać wydatki niestanowiące kosztów uzyskania przychodów

Jestem osobą fizyczną prowadzącą pełną księgowość. Przepisy podatkowe i rachunkowe dotyczące kosztów różnią się od siebie. Czy w rachunku zysków i strat należy wykazać koszty niestanowiące kosztów uzyskania przychodów?

Tak

Ustalając wynik finansowy brutto za dany rok obrotowy należy uwzględnić wszystkie przychody, niezależnie od tego, czy podlegają opodatkowaniu, oraz wszystkie koszty (również rezerwy), nawet jeżeli nie będą stanowić kosztu

sprzedaży walut obcych ogłaszanego przez NBP z dnia zawarcia umowy ubezpieczenia w wartości samochodu przyjętej do celów ubezpieczenia), darowizny, reprezentacja i reklama niepubliczna ponad 0,25 proc. przychodów, wpłaty na PFRON.

Należy również pamiętać o pomniejszeniach lub zwiększeniach przychodów wykazanych w rachunku zysków i strat. **Not. ŁZ**

Czy ewidencjonować udzielenie nieoprocentowanej pożyczki

Spółka zawarła na okres 3 lat umowę pożyczki z jednym ze swoich udziałowców (osobą fizyczną). Zgodnie z zapisami umowy jest to pożyczka nieoprocentowana. W jaki sposób skutki nieoprocentowanej pożyczki udzielonej firmie przez jej udziałowca ująć w księgach rachunkowych spółki?

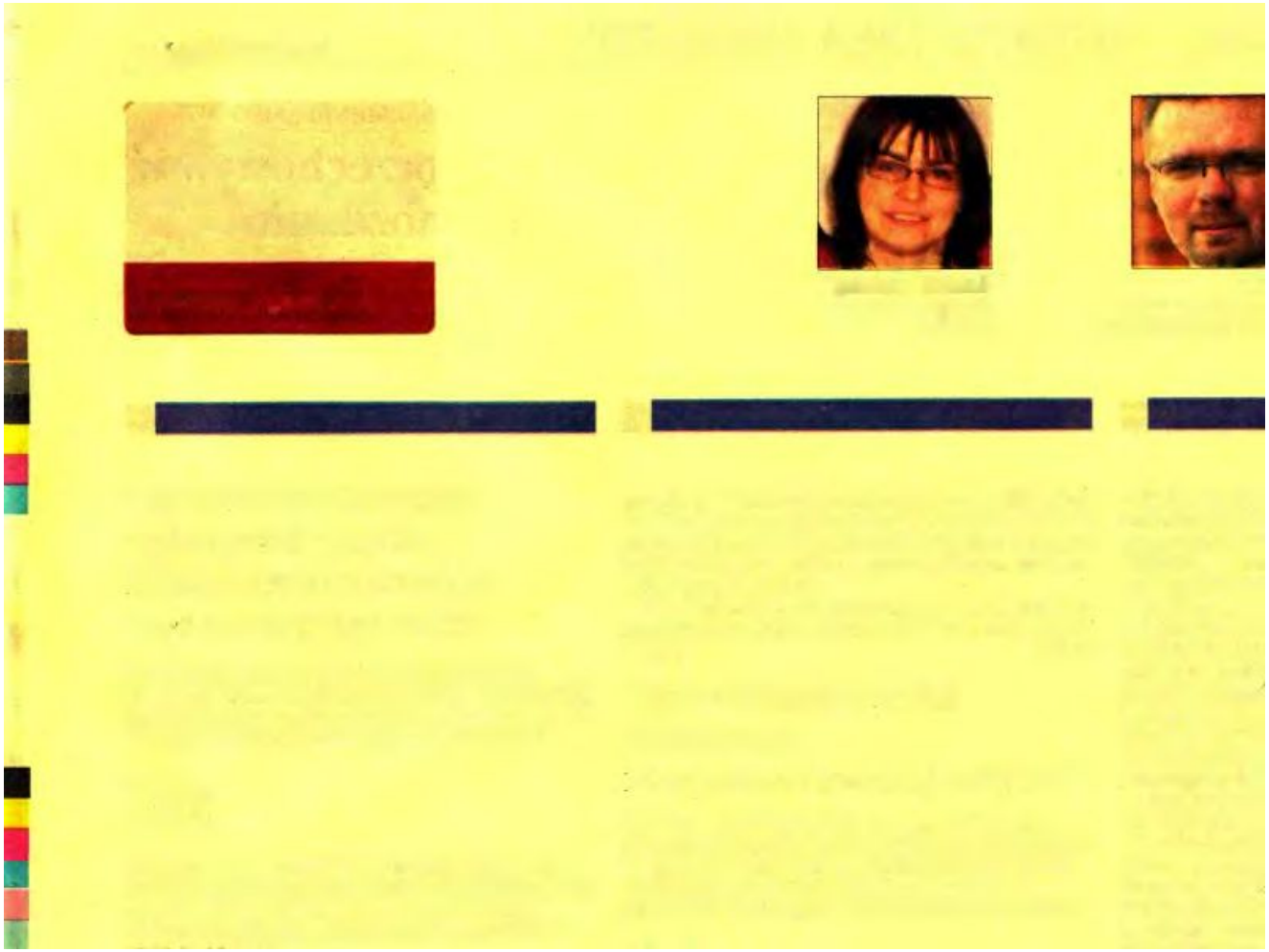
w Krajowym Star dochodowy" (KSI Standardzie Rach (MSR12), na któ ust. 3 ustawy o ra

Według powyż przejściowych po na poziomie spra spółek grupy, różn różnicami pomię westycji w podmi mi pomiędzy jedn zysków lub strat u poszczególnych je

Rozważmy na z wartością inwes

Spółka X posia lidowaną metod kowym sprawozd 100 000 zł. Załó ła nabyta w dniu

Przykład



Przykład

www.gazetaprawna.pl PORADNIA RACHUNKOWA PONIE

Masz pytanie, przyślij e-mail – odpowiemy Ci w dodatku Księgowość i Podatki

ksiegowosc@infor.pl

Agnieszka Michalak
kierownik ds. księgowości w FPA Group

Piot
dyrektor oddziału regio zachód, członek zarządu PKF

Czy w rachunkowości wykazywać wydatki niestanowiące kosztów uzyskania przychodów

Jestem osobą fizyczną prowadzącą pełną księgowość. Przepisy podatkowe i rachunkowe dotyczące kosztów różnią się od siebie. Czy w rachunku zysków i strat należy wykazać koszty niestanowiące kosztów uzyskania przychodów?

Tak

Ustalając wynik finansowy brutto za dany rok obrotowy należy uwzględnić wszystkie przychody, niezależnie od tego, czy podlegają opodatkowaniu, oraz wszystkie koszty (również rezerwy), nawet jeżeli nie będą stanowić kosztu

sprzedaży walut obcych ogłaszanego przez NBP z dnia zawarcia umowy ubezpieczenia w wartości samochodu przyjętej do celów ubezpieczenia), darowizny, reprezentacja i reklama niepubliczna ponad 0,25 proc. przychodów, wpłaty na PFRON.

Należy również pamiętać o pomniejszeniach lub zwiększeniach przychodów wykazanych w rachunku zysków i strat. **Not. ŁZ**

Czy ewidencjonować udzielenie nieoprocentowanej pożyczki

Spółka zawarła na okres 3 lat umowę pożyczki z jednym ze swoich udziałowców (osobą fizyczną). Zgodnie z zapisami umowy jest to pożyczka nieoprocentowana. W jaki sposób skutki nieoprocentowanej pożyczki udzielonej firmie przez jej udziałowca ująć w księgach rachunkowych spółki?

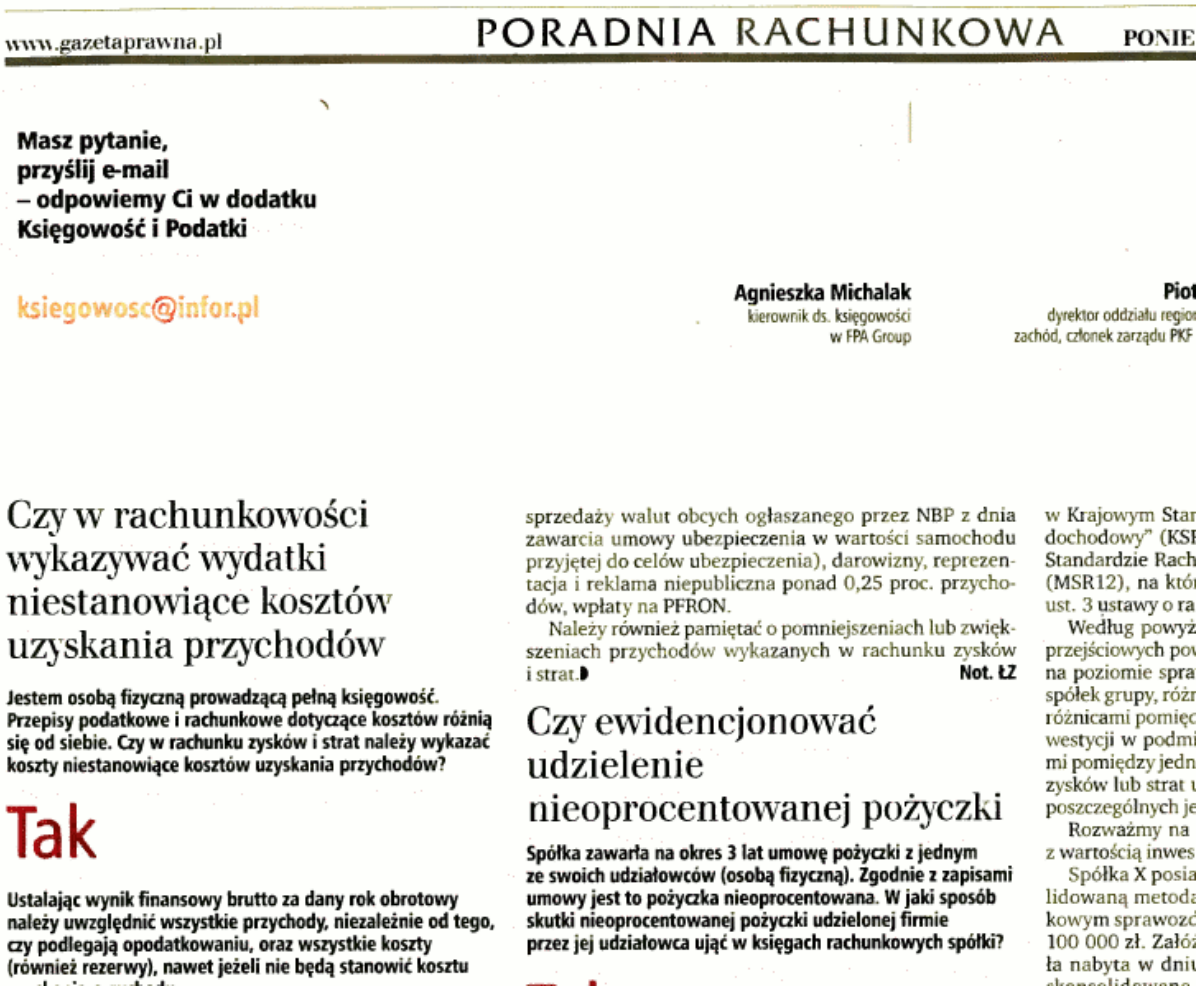
w Krajowym Star dochodowy" (KSI Standardzie Rach (MSR12), na któ ust. 3 ustawy o ra

Według powyż przejściowych po na poziomie spra spółek grupy, różn różnicami pomię westycji w podmi mi pomiędzy jedn zysków lub strat u poszczególnych je

Rozważmy na z wartością inwes

Spółka X posia lidowaną metod kowym sprawozd 100 000 zł. Załó ła nabyta w dniu

Przykład



www.gazetaprawna.pl PORADNIA RACHUNKOWA PONIE

Masz pytanie, przyślij e-mail – odpowiemy Ci w dodatku Księgowość i Podatki

ksiegowosc@infor.pl

Agnieszka Michalak
kierownik ds. księgowości w FPA Group

Piot
dyrektor oddziału regio zachód, członek zarządu PKF

Czy w rachunkowości wykazywać wydatki niestanowiące kosztów uzyskania przychodów

Jestem osobą fizyczną prowadzącą pełną księgowość. Przepisy podatkowe i rachunkowe dotyczące kosztów różnią się od siebie. Czy w rachunku zysków i strat należy wykazać koszty niestanowiące kosztów uzyskania przychodów?

Tak

Ustalając wynik finansowy brutto za dany rok obrotowy należy uwzględnić wszystkie przychody, niezależnie od tego, czy podlegają opodatkowaniu, oraz wszystkie koszty (również rezerwy), nawet jeżeli nie będą stanowić kosztu

sprzedaży walut obcych ogłaszanego przez NBP z dnia zawarcia umowy ubezpieczenia w wartości samochodu przyjętej do celów ubezpieczenia), darowizny, reprezentacja i reklama niepubliczna ponad 0,25 proc. przychodów, wpłaty na PFRON.

Należy również pamiętać o pomniejszeniach lub zwiększeniach przychodów wykazanych w rachunku zysków i strat. **Not. ŁZ**

Czy ewidencjonować udzielenie nieoprocentowanej pożyczki

Spółka zawarła na okres 3 lat umowę pożyczki z jednym ze swoich udziałowców (osobą fizyczną). Zgodnie z zapisami umowy jest to pożyczka nieoprocentowana. W jaki sposób skutki nieoprocentowanej pożyczki udzielonej firmie przez jej udziałowca ująć w księgach rachunkowych spółki?

w Krajowym Star dochodowy" (KSI Standardzie Rach (MSR12), na któ ust. 3 ustawy o ra

Według powyż przejściowych pov na poziomie spra spółek grupy, różn różnicami pomięc westycji w podmi mi pomiędzy jedn zysków lub strat u poszczególnych je

Rozważmy na z wartością inwes

Spółka X posia lidowaną metod kowym sprawozd 100 000 zł. Załó ła nabyta w dniu

Dokument **spakowany** (*bundled multi-page document*):

- jeden plik reprezentuje cały dokument;
- wygodny do przesyłania plików inną drogą niż HTTP;
- czas dostępu do strony: zależny od czasu ściągania poprzednich stron.

Dokument **pośredni** (*indirect multi-page document*):

- główny plik jest tylko indeksem;
- osobny plik na każdą stronę;
- czas dostępu do strony zależny tylko od wielkości tej strony;
- taka sama wygoda przeglądania.

Poza obrazami, dokumenty DjVu mogą zawierać:

- adnotacje:
  - hiperłącza,
  - domyślny sposób wyświetlania,
  - metadane;
- ukryty tekst;
- zakładki;
- miniaturki.

Sposoby udostępniania dokument DjVu na stronach WWW:

- poprzez link do pliku, np.

  `example.djvu?djvuopts&page=42&zoom=100%`;
  zadziała tylko jeśli:
  - serwer zaanonsuje odpowiedni typ MIME (`image/vnd.djvu`) i
  - użytkownik będzie miał zainstalowaną wtyczkę

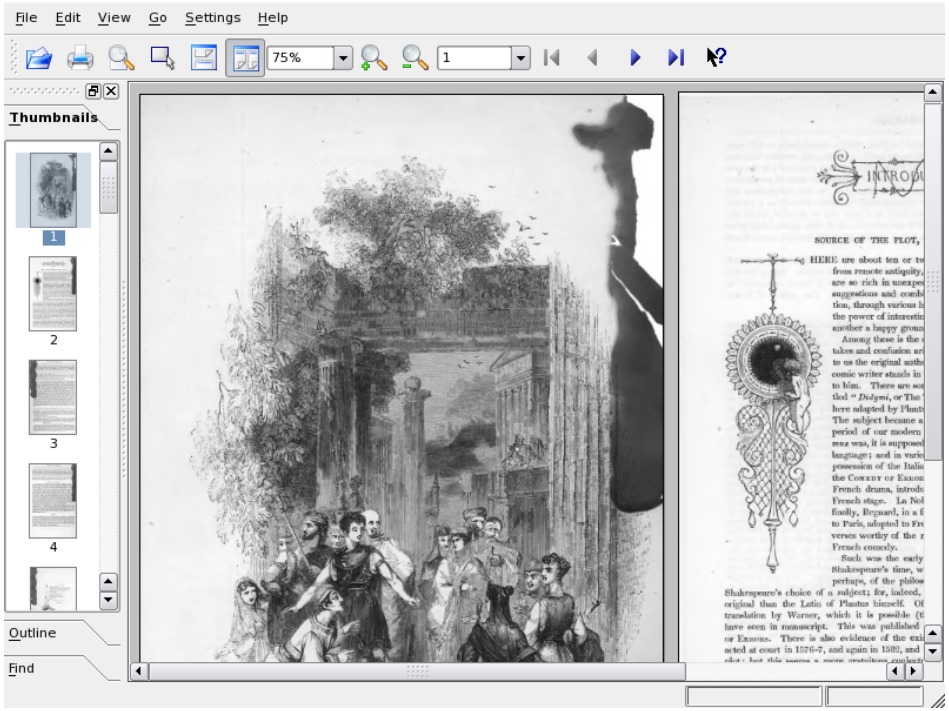
  lub użytkownik będzie wiedział, co zrobić z plikiem po pobraniu;
- osadzenie na stronie HTML, np.

  ```
  <embed src="example.djvu" type="image/vnd.djvu">
  </embed>
  ```
- aplet Javy:
  - ⟨`http://javadjvu.sf.net/`⟩,
  - sprzeczny z zasadą *lekkich* wtyczek,
  - kłopotliwe pobieranie pliku na dysk.

Dostępne oprogramowanie:

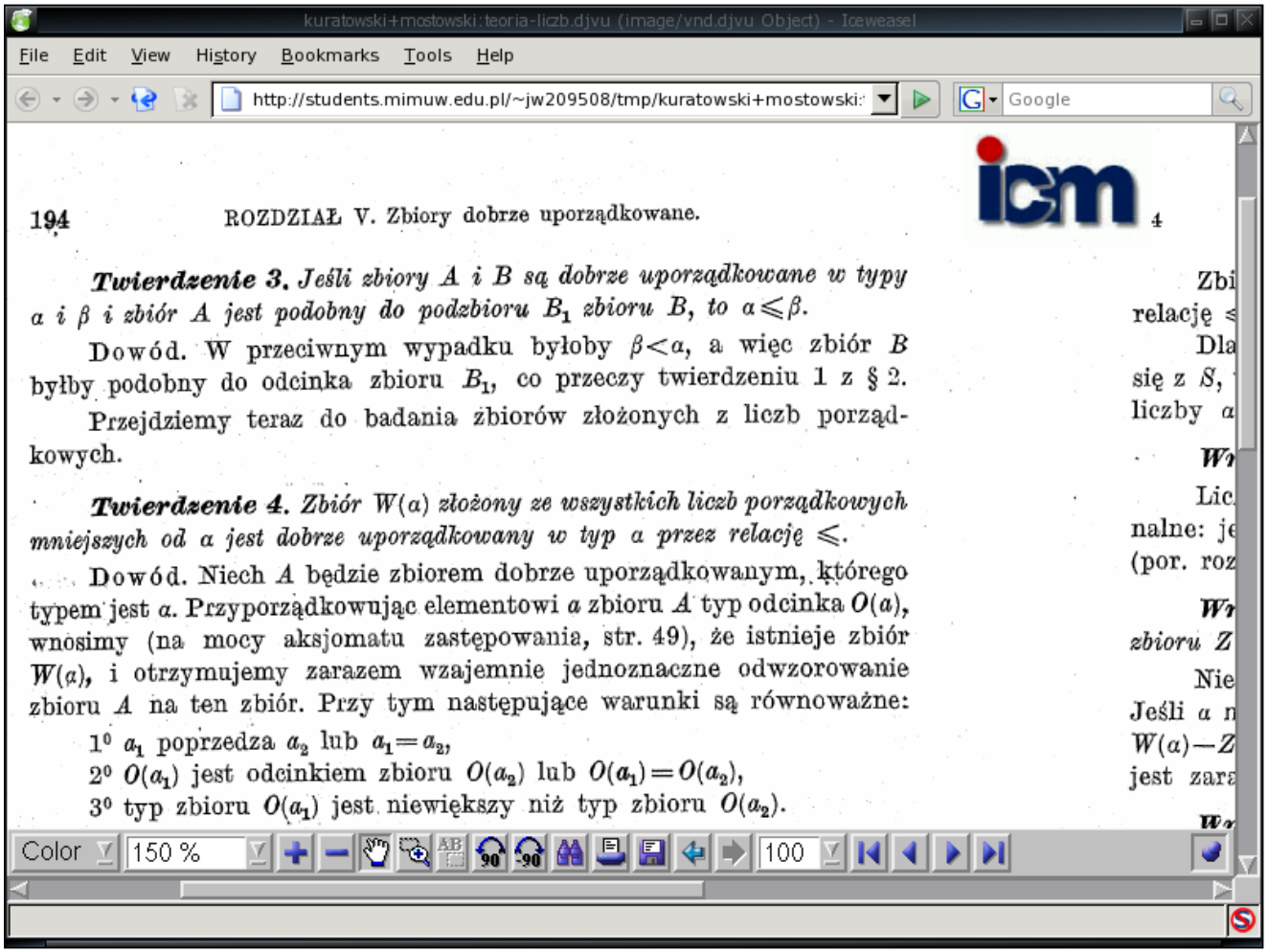
- **DjVuLibre**
  ⟨`http://djvu.sf.net/`⟩
  - na licencji GPL,
  - Linux, inne uniksy, Windows (Cygwin);
- Lizardtech **Document Express**
  ⟨`http://www.lizardtech.com/products/doc/`⟩
  - cena: wysoka,
  - dostępne 30-dniowe wersje testowe,
  - Windows 98/2000/XP lub NT 4.0;
- i inne.

- *djview* (alias *djview3*):
  - siermiężny wygląd;
  - wyświetla jednocześnie co najwyżej jedną stronę.
- *djview4*:



  - korzysta z biblioteki Qt4;
  - brak ograniczeń *djview3*.
- *evince* dla Gnome; *okular* dla KDE.
- *WinDjView* i *MacDjView*.

- *nsdejavu*:



- komercyjne, darmowe wtyczki dla Internet Explorera i Safari
  - ⟨`http://www.lizardtech.com/download/dl_options.php?page=plugins`⟩.

- `c44`:
  - PGM, PPM, JPEG (zdjęcia) → DjVuPhoto;
  - stratna kompresja falkowa;
  - *the encoder requires more memory than necessary*.
- `cjb2`:
  - PBM (bitmapy) → DjVuBitonal;
  - kompresja bezstratna lub stratna;
  - *matching characters on several pages would improve the compression ratios for multi-page documents*.
- `cpaldjvu`:
  - PPM (grafika o małej liczbie kolorów) → DjVuLayered;
  - brak możliwości wyboru koloru tła;
  - brak kontroli nad kwantyzacją kolorów.
- `csepdjvu`:
  - PPM + własne formaty RLE → DjVuLayered;
  - raczej do użytku przez inny program;
  - potrafi sprytnie włączyć do dokumentu warstwę tekstową.

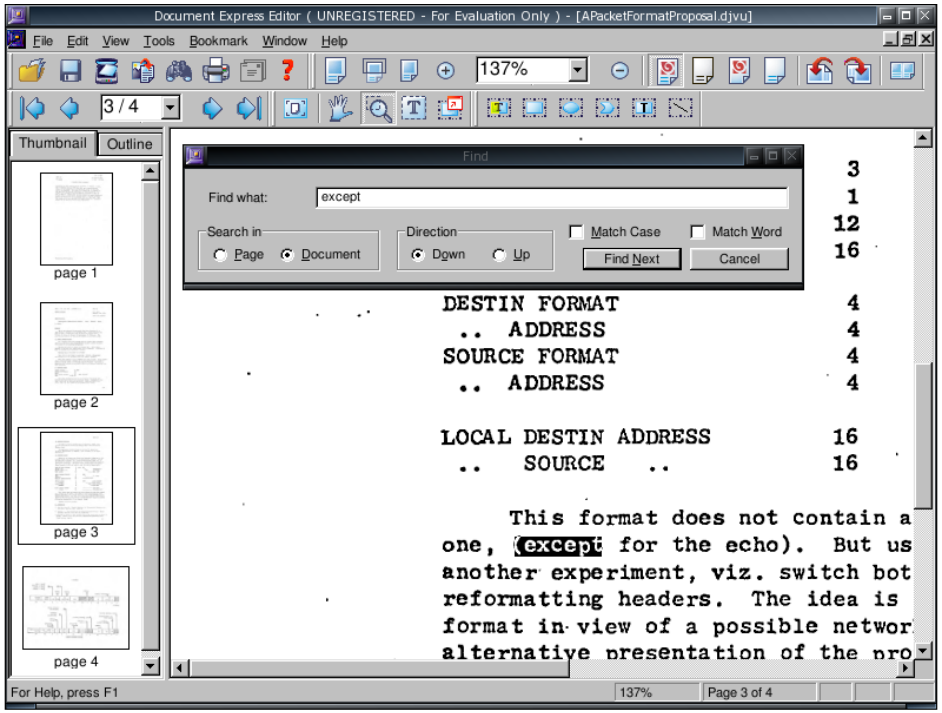
- `djvudigital`:
  - PostScript, PDF → DjVu;
  - wymaga specjalnego sterownika dla *Ghostscripta*, którego nie można dystrybuować w formie binarnej;
  - opcjonalnie – włącza do dokumentu tekst;
  - nie potrafi zachowywać hiperłączy, zakładek ani metadanych.
- `any2djvu`:
  - DjVu, PostScript, PDF, TIFF, JPEG, PNM i inne → DjVu;
  - korzysta z serwisu online;
  - opcjonalnie – wykonuje OCR;
  - w przypadku PDF: zachowuje hiperłącza;
  - w przypadku PDF/PostScript: nie potrafi zachowywać tekstu, zakładek ani metadanych.

- `djvumake`:
  - montuje jedną stronę dokumentu DjVu z kawałków;
  - nie pozwala włączyć kawałka z adnotacjami.
- `djvm`:
  - łączy dokumenty DjVu w spakowany dokument;
  - wstawia dokument DjVu do spakowanego DjVu;
  - usuwa stronę ze spakowanego dokumentu.
- `djvmcvt`:
  - spakowany DjVu $\leftrightarrows$ DjVu pośredni.

- `djvused`:
  - tworzy/edytuje/usuwa:
    - adnotacje,
    - ukryty tekst,
    - zakładki;
  - generuje/usuwa miniaturki stron;
  - konwertuje: spakowany DjVu $\leftrightarrows$ DjVu pośredni.
- `djvutoxml` i `djvuxmlparser`:
  - adnotacje i ukryty tekst $\leftrightarrows$ XML;
  - nie eksportuje/importuje zakładek ani miniaturek;
  - nieprzyzwoicie niewydajny.
- `EmbedBookmarks`:
  - tworzy zakładki na podstawie pliku HTML o prostej strukturze;
  - ⟨`http://windjview.sourceforge.net/bookmarks.html`⟩.

- `ddjvu`:
  - DjVu $\rightarrow$ PPM, PGM, PBM, TIFF, własny RLE;
  - opcjonalnie – wizualizacja tylko fragmentu strony (stron).
- `djvups`:
  - DjVu $\rightarrow$ (Encapsulated) PostScript;
  - opcjonalnie – tryb książeczki.
- `djvuextract`:
  - wyodrębnia kawałki z dokumentu DjVu.
- `djvutxt`:
  - wyodrębnia tekst osadzony w dokumencie DjVu.

- *Document Express Professional*:
  - *Document Express Editor*:



    - graficzny edytor dokumentów DjVu,
    - manipulowanie ukrytym tekstem — tylko OCR (ReadIris),
    - działa pod winem;
  - *Virtual Printer Pro*:
    - wirtualna drukarka – tworzenie DjVu z dowolnego programu obsługującego drukowanie;
- cena (wersja 6.0): 1,35 tys. zł + VAT.

- *Document Express Enterprise*:
  - możliwości *Document Express Professional* oraz:
    - narzędzia do wsadowej konwersji dokumentów do DjVu,
    - tryb *hot folders*;
  - wymaga *.NET Framework* — nie działa pod winem;
  - cena: (wersja 5.1) 23 tys. zł + VAT.
- *DjVu Solo*:
  - graficzny edytor dokumentów DjVu;
  - nie można edytować ukrytego tekstu ani zakładek;
  - działa pod winem;
  - za darmo, ale tylko do zastosowań niekomercyjnych.

pdf2djvu

⟨`http://freshmeat.net/projects/pdf2djvu/`⟩

- PDF → DjVu;
- autor: Jakub Wilk;
- licencja: GPL 2;
- włącza do dokumentu:
  - warstwę graficzną:
    - pierwszy plan: tekst, grafika wektorowa, grafika rastrowa bitonalna,
    - tło: reszta,
  - tekst,
  - hiperłącza,
  - zakładki,
  - metadane;
- użyte narzędzia:
  - biblioteka *poppler*,
  - biblioteka *DjVuLibre*:
    - publiczne API,
    - `csepdjvu`, `djvuextract`, `djvused`, `djvumake`, `djvm`.

- Programy narzędziowe:
  - niekonsekwentne nazewnictwo;
  - niekonsekwentne ograniczenia;
  - proste problemy wymagają nieintuicyjnych zabiegów;
  - nie zawsze są prostymi opakowaniami na funkcje biblioteczne.
- API publiczne:
  - dla C i C++;
  - skromny zakres:
    - obsługa S-wyrażeń[2],
    - dekodowanie DjVu;
  - asynchroniczna natura;
  - brak zależności od protokołów sieciowych.
- API prywatne:
  - tylko dla C++,
  - dokumentacja sprzeczna z rzeczywistością!
  - niestabilne?

- Programy narzędziowe:
  - niekonsekwentne nazewnictwo;
  - niekonsekwentne ograniczenia;
  - proste problemy wymagają nieintuicyjnych zabiegów;
  - nie zawsze są prostymi opakowaniami na funkcje biblioteczne.
- API publiczne:
  - dla C i C++;
  - skromny zakres:
    - obsługa S-wyrażeń[2],
    - dekodowanie DjVu;
  - asynchroniczna natura;
  - brak zależności od protokołów sieciowych.
- API prywatne:
  - tylko dla C++,
  - dokumentacja sprzeczna z rzeczywistością!
  - niestabilne?

[2] *Any sufficiently complicated C or Fortran program contains an ad-hoc, informally-specified bug-ridden slow implementation of half of Common Lisp.* /Philip Greenspun/

Do zrobienia:

- narzędzie do efektywnej konwersji: ukryte dane $\leftrightarrows$ XML;
- dalsza integracja DjVu — OCR;
- bindingi biblioteki *DjVuLibre* dla Pythona:
  - API publiczne API,
  - API prywatne?;
- graficzny edytor DjVu.

- DjVu Zone
  ⟨http://www.djvuzone.org/⟩
- *DjVu Technology Primer*
  ⟨http://www.lizardtech.com/files/doc/techinfo/DjVu_Tech_Primer.djvu⟩
- *Overview of the DjVu Document Compression Technlogy*
  ⟨http://www.lizardtech.com/files/doc/techinfo/2001_compression_overview.djvu⟩
- Léon Bottou *High Quality Document Image Compression with DjVu*
  ⟨http://leon.bottou.org/slides/djvu/index.djvu⟩
- *Lizardtech DjVu Reference*
  ⟨http://www.lizardtech.com/files/doc/techinfo/DjVu3Spec.djvu⟩