

Настанова з творення djvu-книжок.

Цю настанову побудовано з використанням матеріалів сайту

<http://www.djvu-soft.narod.ru/>

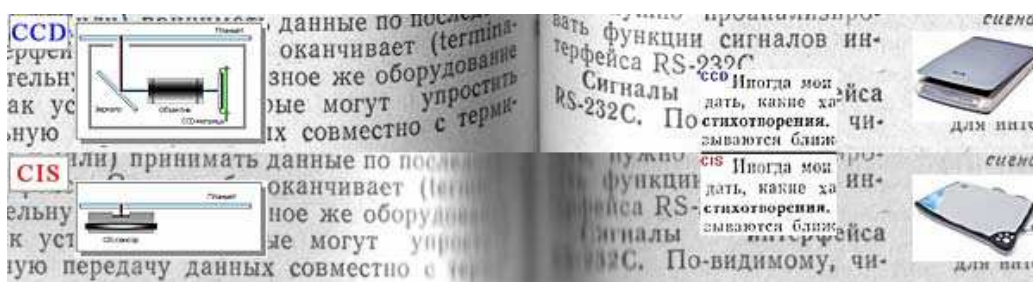
та власного досвіду.

Прелюдія. Якого треба сканера

Почнемо з сканера. Купуючи сканера, треба звернути увагу на підтримувану роздільну здатність та систему. Зараз представлено дві системи сканерів

1. CIS (Contact Image Sensor, контактовий давач образу).
2. CCD (Charge-Coupled Device, прилад із зарядозв'язком).

На фото показано порівняння тексту засканованого сканерами різних систем.



Як бачимо, якість сканування CCD-сканера значно краща. Тому саме такий сканер і треба купувати. Ціни сканерів майже однакові.

Докладніше про CIS та CCD сканери можна тут (стаття 2007 року).

http://www.infanata.org/2007/07/24/nagljadnoe_sravnenie_skanerov_ccd_i_cis_pri_skanirovanii_knig.html

Якщо Ви будете цифрувати книжки у дуже великих кількостях, радимо придбати спеціалізованого книжкового сканера Plustek OpticBook 3600 або Plustek OpticBook 4600. Його в постійному продажу немає, але можна замовити. Він займає проміжне положення між звичайними офісними сканерами та дуже дорогими спеціалізованими книжковими сканерами.



Цей сканер досить дорогий, але на ньому не псуються книжки (не треба притискати розгорнуту книжку до скла сканера) і він має велику швидкість сканування (4 секунди сторінка в градаціях сірого і роздільною здатністю 300 dpi).

Сканування й творення djvu-книжок.

Деякі постулати (поради досвідчених сканувальників).

Наводимо їх без пояснень.

1. Сканувати треба з **роздільною здатністю 300 dpi у градаціях сірого (greyscale)**.
2. **Використовувати** для збереження отриманих сканів лише формат **tiff без стиснення**.
3. **Не використовувати** для сканування програму FineReader, за винятком того випадку, коли Ви хочете створити текстовий варіант книжки.
4. Для книжок-образів використовуйте лише формат djvu. Pdf використовуйте лише для текстових книжок.
5. **Якщо на скані прозирає текст чи ілюстрації зі зворотнього боку сторінки, підкладіть під сторінку чорний папір.**

Далі за етапами Технології виготовлення djvu-книжки.

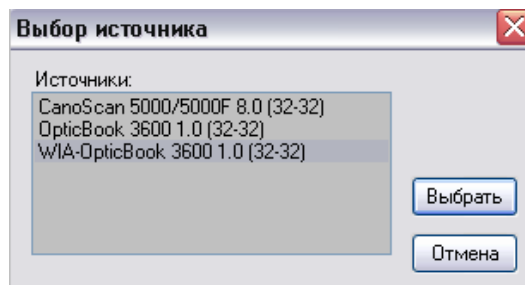
1. Сканування. Отримання необроблених („сирих”) образів.
2. Оброблення „сирих” сканів. Отримання робочих образів.
3. Кодування. Отримання djvu-файлу книжки.
4. Створення OCR-шару.
5. Створення навігації.

1. Сканування.

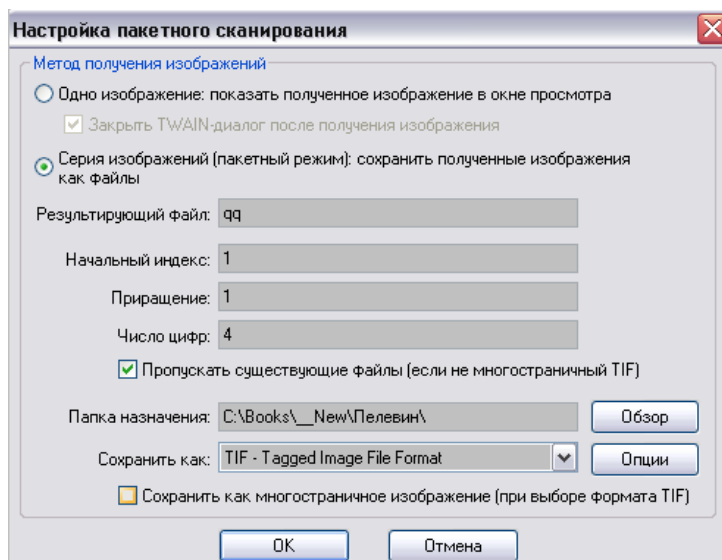
Сканувати можна використовуючи будь-яку програму, що дозволяє керувати сканером і зберігати отримані зображення у файлах з послідовною нумерацією.

Далі подаю порядок роботи з безплатною програмою **IrfanView**.

У меню ФАЙЛ натискаємо пункт **Выбрать TWAIN-источник**.

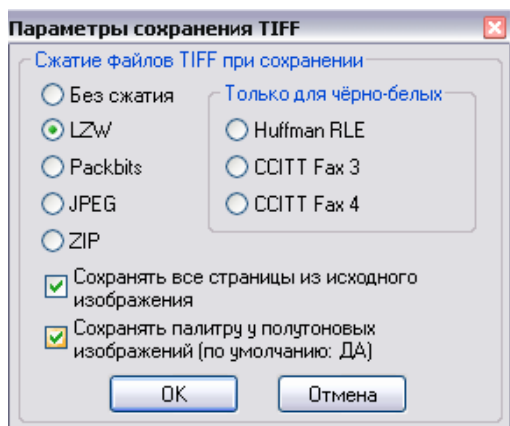


Далі у тому ж меню вибираємо пункт **Получить изображение/пакетное сканирование...**



Обов'язково вибираємо формат tif, бо цей формат не розмиває зображення і не допускає втрати якості образу. Задаємо шлях до теки, де буде збережено скани. Краще зберігати скани в окремих файлах, а не в багатосторінковому файлі. Це дає змогу легко пересканувати невдалий образ. Загальну частину назви (пункт **Результіруючий файл**) треба робити невеликою, що запобігти створенню довгих назв. Назву книжки краще дати тежчї де буде збережено скани. Параметри, встановлені в інших пунктах, задовільні для будь якої книжки з меншою за 9999 кількістю сторінок.

Обов'язково треба перевірити **Опции** графічного формату



Тут найкраще обрати **Без сжатия**, оскільки не всі програми коректно працюють з стисненими файлами. Тиснемо **ОК** і починаємо сканування.

Пакетне сканування виконуємо в слїпу, що дає максимальну швидкість сканування. Невеликих перекосів не треба лякатися, бо їх буде виправлено далї.

Отже, етапи сканування такі:

Кладемо розгорнену книжку на скло сканера і притискаємо корїнець рукою. Притискання кришкою сканера або вантажами значно довшї.

Робимо попереднє сканування (**Preview**) і встановлюємо зону сканування. Треба мати невеликий запас за меншою стороною робочого поля сканера, аби книжка завжди потрапляла до робочої зони.

Починаємо сканування. Перегортати сторїнки книжки можна вже на зворотньому ходї каретки сканера.

Треба стежити, аби книжка не виходила за межї робочої зони і мїцнїше (але не з усїєї сили) притискати корїнець, аби геометричнї спотворення були якомога меншими.

На виходї маємо файли „сирих” образів у форматї **tif з роздїльною здатнїстю 300dpi у градацїях сїрого**. Розмїр кожного файлу близько 8 мегабайт.

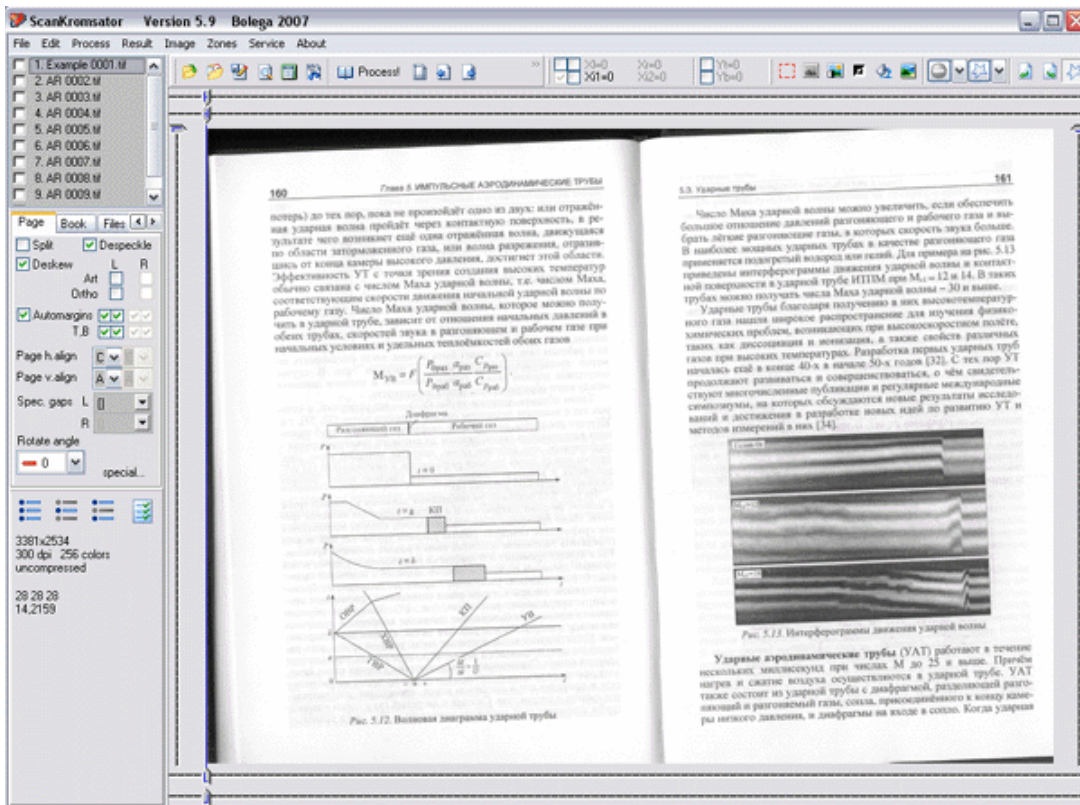
Швидкїсь сканування може досягати 200 розгорток за годину, це книжка на 400 сторїнок. Тепер можна перейти до обробляння „сирих” сканїв.

2. Обробляння сканїв.

Обробляти отрманї скани можна рїзними програмами, але тут буде описано безплатну програму **ScanKromsator**.

Це дуже потужна програма обробляння сканїв з неочевидними для новачка властивостями. Тому наводимо покрокову їнструкцїю.

Запускаємо програму і завантажуюємо в неї наші файли.



У верхньому лівому кутку розташовано список завантажених файлів. Під списком розташовано закладки, що ними ми зараз і займемося.

Закладка **Файл**.

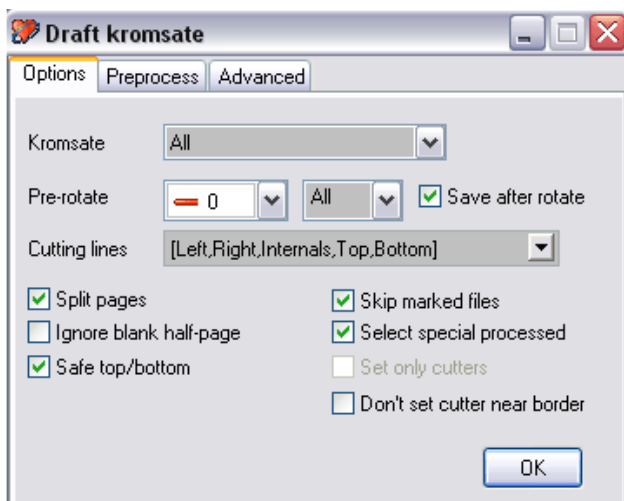
Тут ми задаємо шлях для результатів краєння („кромсання”) та обов’язково **призначаємо вихідну роздільну здатність 600 dpi**. Тут же можна задати спосіб нумерації виходових файлів.

Робимо чернеткове краєння.

Лівіше від кнопки **Process** розташовано кнопку з ножицями (**Draft kromsate**).

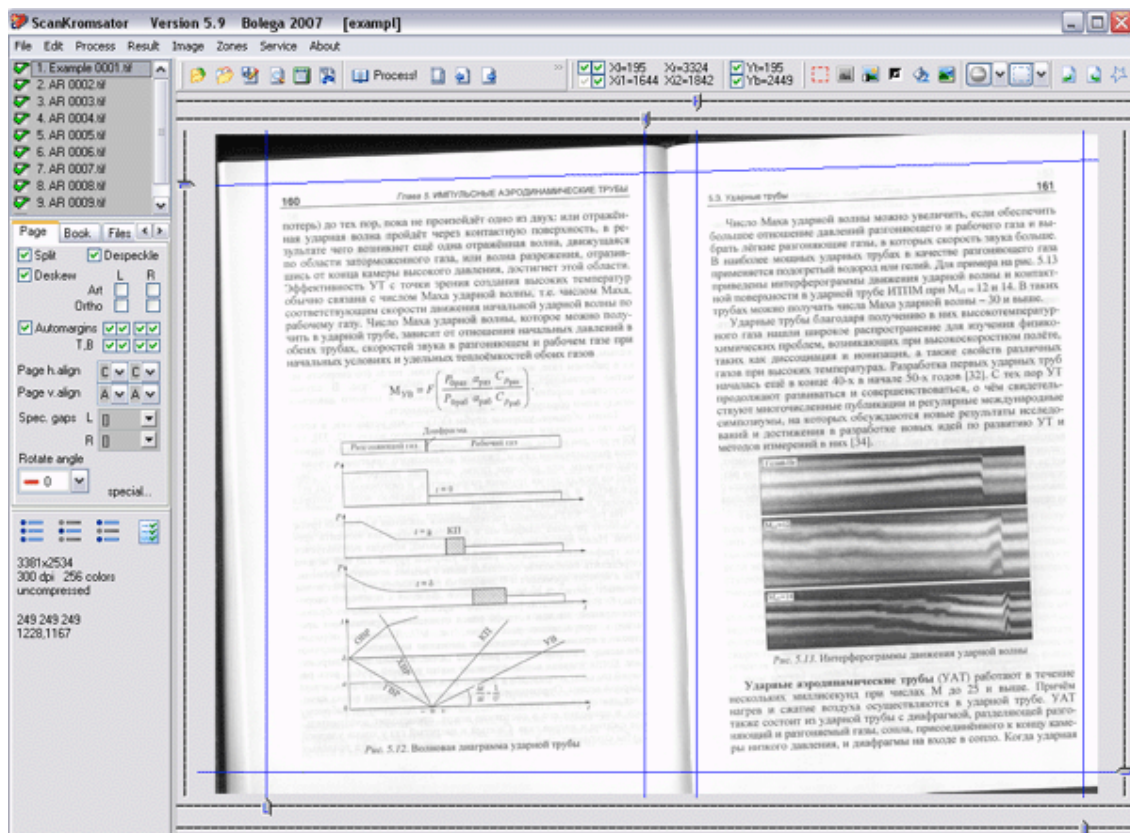


Натискаємо на неї й отримуємо вікно діалогу.



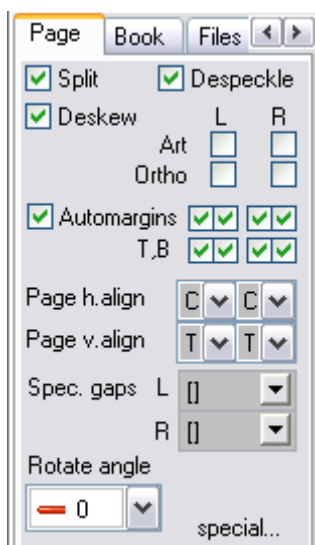
Ставимо пташки на **Split pages** і **Safe top/bottom** і тиснемо **OK**.

Через деякий час, що залежить від швидкодії Вашого комп'ютера та обсягу книжки (в середньому це близько 10 хвилин), отримуємо результати. Сині лінії — то лінії різання: усе, що поза ними, буде викинуто. Також з'явилися зелені пташки біля назв файлів. Це означає, що програма їх чернетково обробила.



Центральні лінії розрізу показують, що сторінку буде поділено на дві окремі, а центральну частину буде відрізано.

Настав дуже важливий етап — розставити опції. Це роблять у закладках під списком файлів. Якщо опцію застосовують до сторінки, а її треба застосувати до усіх сторінок, то вибираючи значення утримують клавішу **Ctrl**.



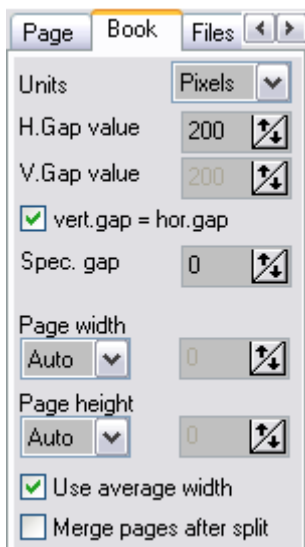
Pages.

Тут виставляємо спосіб центрування. Усталено тут стоїть **A** — автомат. Це означає, що зображення буде розміщено у верхньому лівому кутку. Але, залежно від форматування книжки, можна поставити й інші значення для горизонтального (**Page h.align**) та вертикального (**Page v.align**) вирівнювання. Тут **B** — вирівнювання вниз, **T** — вгору, а **C** — центрування.

Despeckle – прибирання дрібних плямок.

Deskew – вирівнювання нахилу сторінок. Погано вирівняну сторінку можна переробити методом **Art**. Застосування його до всіх сторінок істотно сповільнює процес. Для сторінок, де текст повернуто на 90 градусів, застосовуємо метод **Ortho**. Завважимо, що ці методи задають окремо для лівої (L) та правої (R) частин

розгортки.



Book.

Тут виставляють розміри виходових сторінок. **Page width** та **height** залишаємо усталене значення **Auto**. **H.Gap value** задає розміри берегів. Для 600 dpi це зазвичай 200 або 250 pixels, але можна вибрати й інші розміри.

Files.

Повинна стояти роздільна здатність **600 dpi**. Це надзвичайно важливо й від цього залежить результат.



Options.

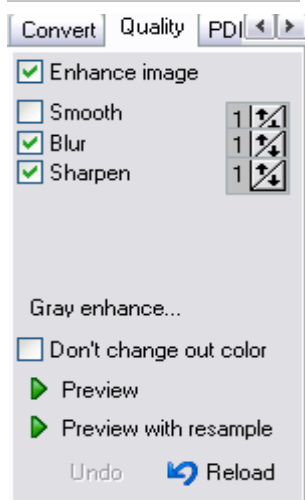
Ставимо **Deskew method = Auto(shear)**. **Despeckle**: задаємо або метод **Safe** або **Fine+Normal**. Другий – це інтелектуальний метод чищення, він, наприклад не вичищає точки над і. Можна також пересунути на дві три поділки повзунки **Text sensitivity** щоб не було відрізано відокремлені від тексту номери сторінок.

Options 2. Пропускаємо.



Convert.

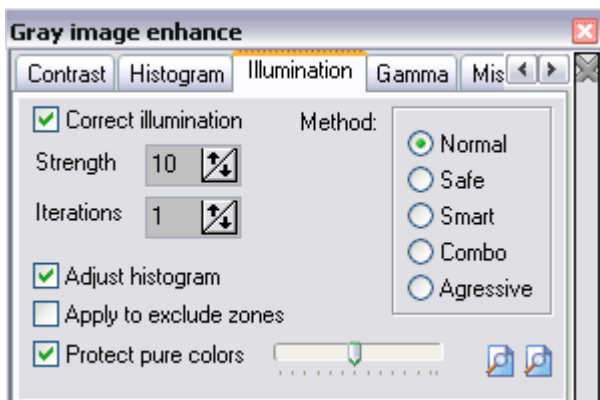
Тут виставляємо поріг для перетворювання градацій сірого у чорно-білий. Обираємо **MiddleDark**. При цьому утримуємо клавішу **Ctrl**, аби застосувати опцію до всіх сторінок.



Quality.

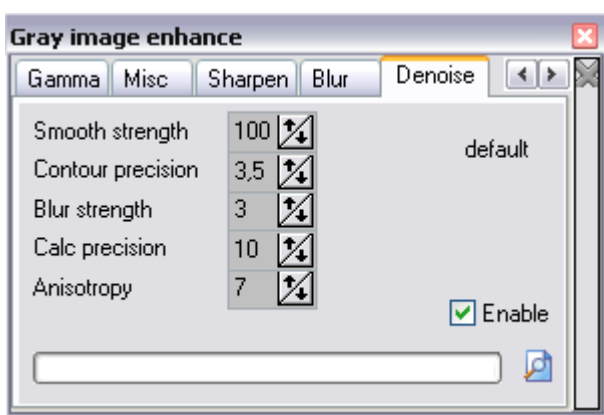
Дуже важлива опція. В **Enhance image** позначаємо пташками **Blur** и **Sharpen**. Їхні значення залежать від шрифту засканованої книжки. Звичайно вони дорівнюють 1 чи 2. Підвищення значення призводить до потовщення ліній, що утворюють літеру.

Дуже **важливо!** Якщо скани у градаціях сірого, то тиснемо на **Gray enhance**. Це викликає діалог **Gray image enhance**.



У викликаному діалозі, обираємо закладку **Illumination**, де ставимо пташку на **Correct illumination**.

За цієї опції програма вирівнює освітленість образу, що прибирає чорні смуги і багато сміття. Особливо важливо для центру розгортки.



У закладці **Denoise** ставимо пташку на **Enable**. Параметри — як на рисунку.

Усі параметри для крайня виставлено. Аби щоразу не повторювати цю процедур, можна створити свій профіль **File->Options...**

Тепер треба перевірити правильність розташування ліній розрізу та виділення ілюстрацій (якщо вони є) на всіх сторінках.

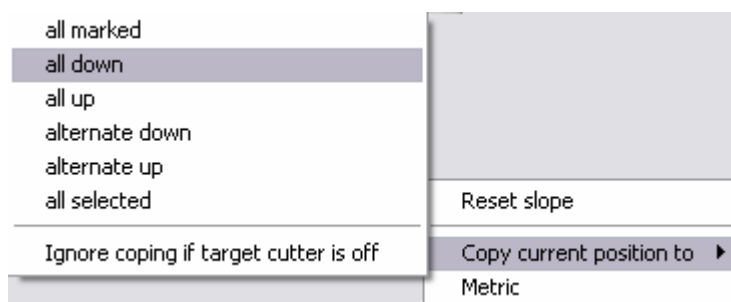
Якщо на якійсь сторінці різак розташовано неправильно, посуваємо їх у правильне положення. Там, де треба, змінюємо на закладці **Pages** спосіб центрування сторінки, а в разі сторінок з текстом, повернутим на 90 градусів, встановлюємо **Deskew =Ortho**.

Буває так, що сторінку розташовано під кутом, або тінь на розгортці ширшає. Тоді можна встановити скісні лінії розрізу. Отже, пересуваємо різак за кінчик за натисненої клавіші **Shift**.

Фахівці радять на цьому етапі такий оптимальний алгоритм:

Лівою рукою гортаємо (клавіші **q** та **w**), права рука на мищі, аби за потреби пересунути різак.

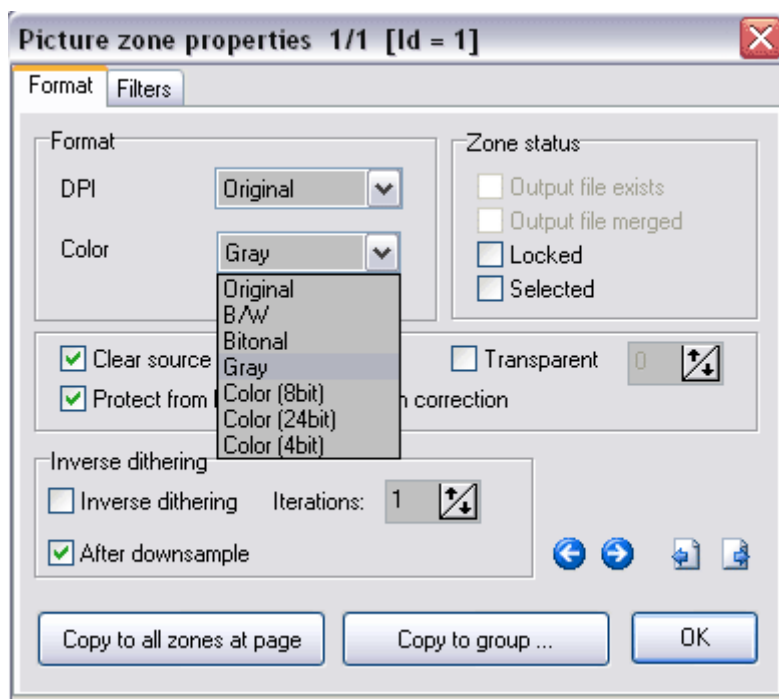
Якщо для частини сторінок розташування різаків повинно бути однаковим, то можна скопіювати їхнє розташування. Для цього треба натиснути на різак праву клавішу мишки і вибрати потрібну опцію (**Copy current position to**).



Якщо на сторінці є фото (чорно-біле чи кольорове) або напівтонова ілюстрація, то їх виділяють у **Picture Zones** і для них передбачено спеціальний режим оброблення. Ілюстрацію виділяємо мишкою прямокутником і натискаємо кнопку **Mark as picture zone**. Для скісно засканованих сторінок чи ілюстрацій непрямокутної форми можна використати **Polygon selection**. На відповідній кнопці зображено перехняблену зірочку.



В результаті оброблення усі виокремлені ілюстрації буде збережено в окремих файлах.



Параметри **Picture zone** можна налагодити. Подвійний клац на виділеній зоні викликає діалог налаштування **Picture zone properties**, де треба для кольорових ілюстрацій виставити параметра **Color**, усталено виставлено **Gray**.

Оскільки ілюстрації виділені в окремі файли, нам треба їх об'єднати зі сторінками книжки. Обираємо пункт меню **Zones->Picture Zone->Merge zones...** і це все, файли об'єднано. Для творення книжок з високоякісними ілюстраціями є спеціальні програми, але це окрема пісня.

Завдання можна зберегти **File->Save Task**

Усе. Починаємо сам процес краєння — натискаємо клавішу **Process**. Програма запитує, чи ми справді хочемо змінити роздільну здатність, і їй можна сміливо сказати ОК, бо саме заради цього ми проробили все описане.



Тепер працює комп'ютер..

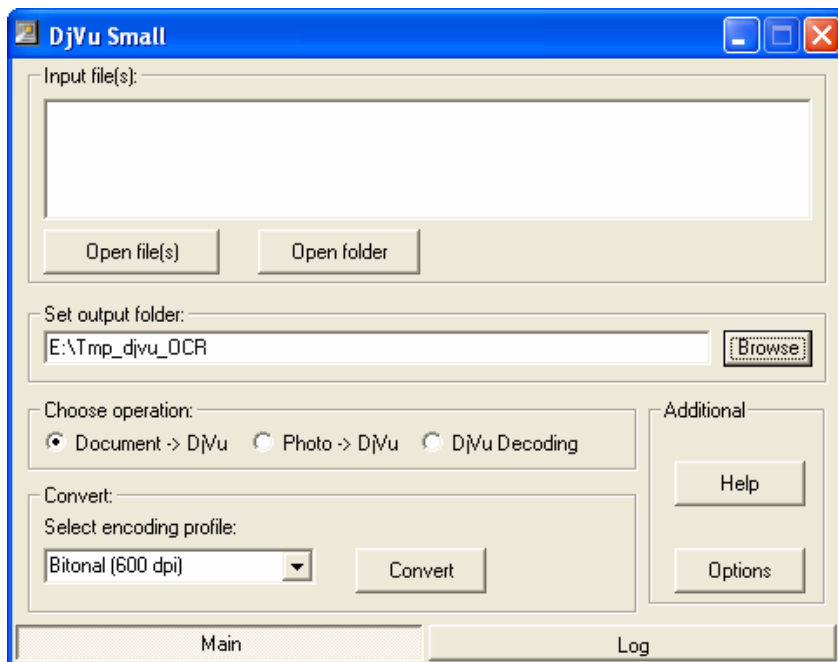
Приблизно за 20 хвилин краєння закінчене і результати збережено у призначеній течці.

Уважно передивляємося результати. Можуть бути неправильно вирівняні образи, виправляємо їх окремо. Можна додатково почистити отримані образи. На це є досить потужний набір інструментів.

3. Кодування

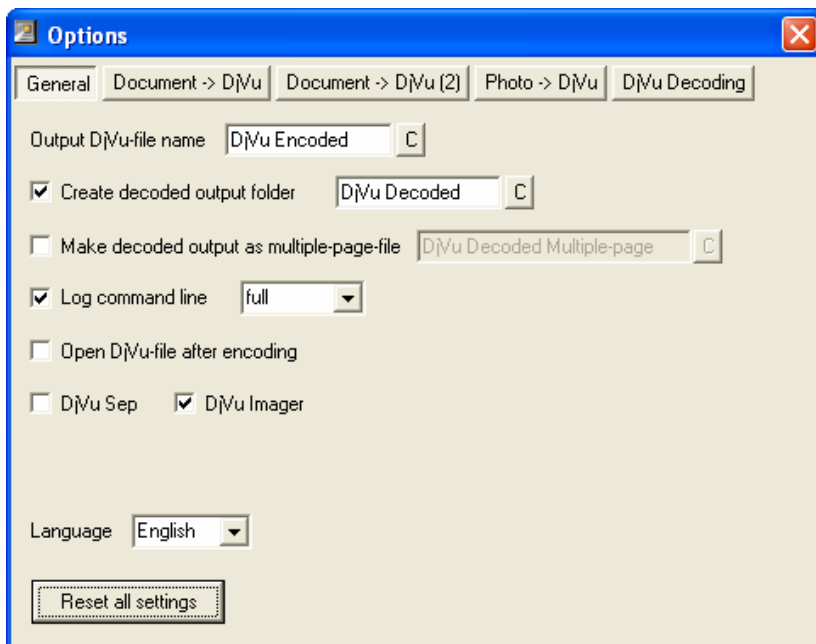
Перетворити отримані образи в djvu можна кількома шляхами. Найпростіший — використати програму **DjVu Small**. Це дуже легкий (менше 300 кілобайтів) але потужний застосунок. Я не буду казати про деякі додаткові можливості цієї програми. Далі — лише мінімальні настанови.

Запускаємо програму й отримуємо таке вікно.

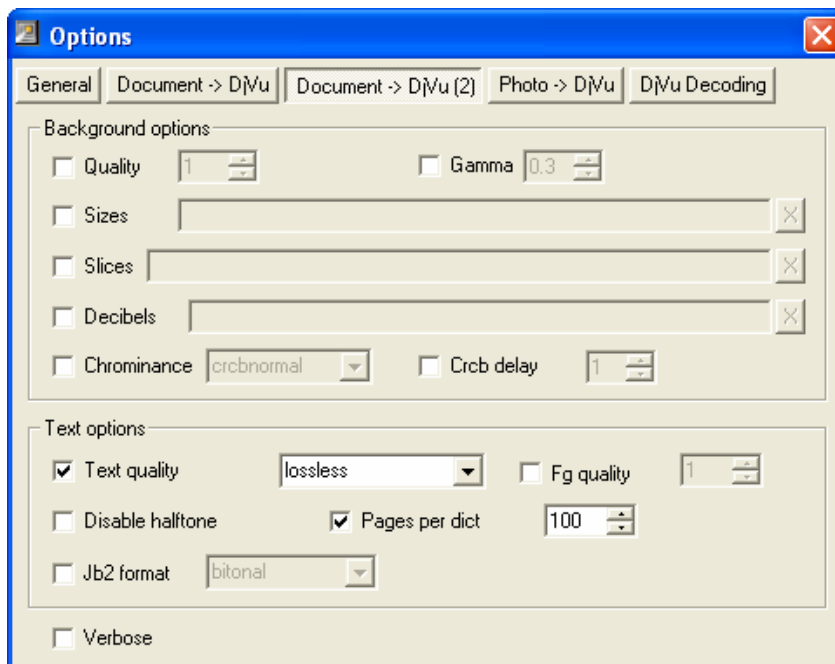


У ньому встановлюємо ім'я файлу в разі поодинокого образу чи багатосторінкового tif-файлу, або відкриваємо теку з обробленими файлами образів. Встановлюємо виходову теку. Далі вибираємо профіль кодування. Для книжок без напівтонових ілюстрацій обираємо Vitonal (600 dpi). Якщо у книжці є фото чи напівтонові ілюстрації обираємо Scaned (600 dpi).

Один раз треба встановити опції. Натискаємо клавішу **Options** і отримуємо вікно діалогу.



Тут залишаємо усталені значення. Зміни вносимо у закладці **Document->DjVu(2)**. У розділі **Text options** ставимо пташку на **Text quality** і обираємо **lossles**.



Далі ставимо пташку на **Pages per dict** та ставимо число, від 100 до 1000 (останнє— це вияв параноїдного екстремізму). Така опція має скоротити обсяг файлу до 25%. Але ці значення опцій не викликають загального одобрямсу. Є думка, що число сторінок на словник не повинно перевищувати 20. То експериментуйте.

Ці опції зберігаються надалі і їх не треба встановлювати знову.

Тепер тиснемо клавішу **Convert** і чекаємо результатів. У перебігу бачимо градусника, що показує виконання кожної окремої сторінки.

Нарешті отримуємо повідомлення, що конвертацію успішно завершено.

В результаті в заданій течці з'являється файл DjVu Encoded.djvu. Переназиваємо файла.

Ми отримали файл образів у форматі djvu. Тепер нам треба додати OCR-шар та навігацію (якщо це потрібно).

4. Творення текстового шару.

На це потрібні дві програми:

1. **FineReader 7.0** або **8.0** версії
2. **DjvuOCR 2.2**

Перша з них комерційна, але досвідчені люди кажуть, що для наших справ досить тріальної версії **FineReader**, що її можна вільно завантажити з фірмового сайту.

Друга програма безплатна, її створив програмувальник з Болгарії **Gencho**.

Крок перший.

Завантажуємо у **FineReader** отримані на **ScanKromsator**'і файли і запускаємо читання пакету. В старих версіях програми **DjvuOCR** (до **2.1** включно) можливості редагувати отриманий OCR текст не було. Починаючи від версії **2.2**, така змога постала, але з певними обмеженнями:

1. Під час редагування зберігати деякі символи оригінального тексту. Тобто не викидати великі блоки тексту.
2. Зберігати кількість рядків. Тобто не стирати та не додавати символів кінця рядка.

Усі кадри повинні бути або розпізнано, або (коли цей кадр не треба текстувати) виділено як образ.

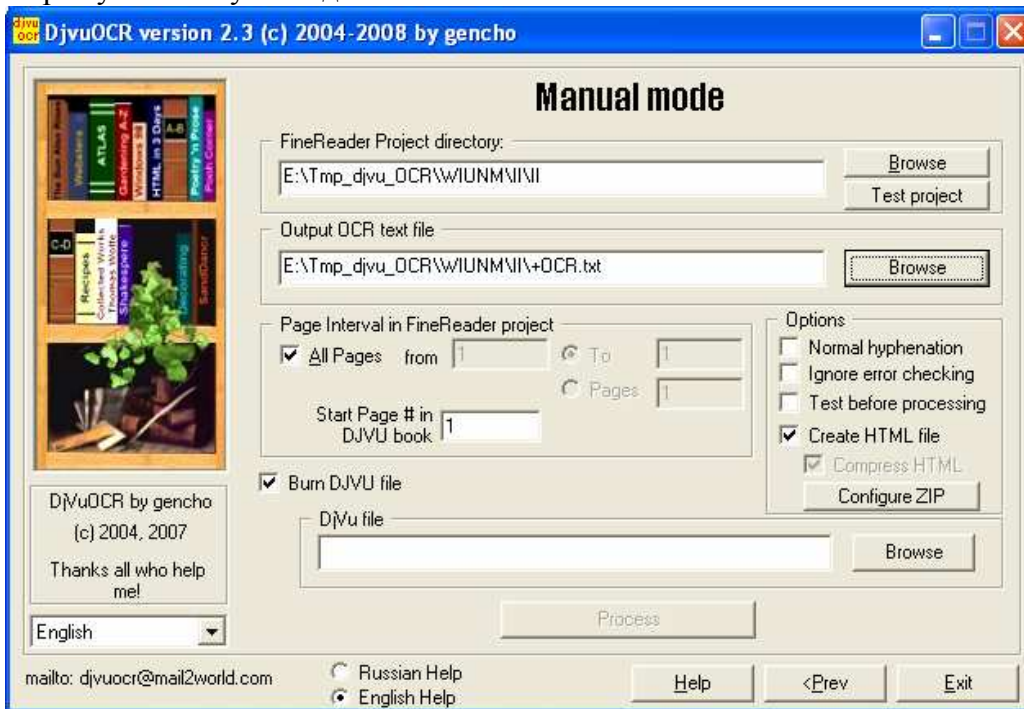
Вичитуємо тест і робимо

Крок другий.

Запускаємо **DjvuOCR** і натискаємо клавішу **Manual made OCR manager**.



Отримуємо наступний діалог



У ньому:

- У **FineReader Project directory** позначаємо теку проекту **FineReader**.
- У **Output OCR text file** подаємо назву будь-якого порожнього файлу в течці проекту.
- Ставимо пташку біля **Burn DJVU file**.
- Вибираємо djvu-файл з книжкою в **DJVU file**.
- Тиснемо клавішу **Process**.

Чекаємо декілька хвилин.

Це все. Наша книжка має текстовий шар.

5. Творення навігації.

Є кілька способів задати навігацію в книжці.

Існує навіть утиліта для творення навігації за змістом книжки **DjVu Hyperlinks Editor**. Однак її розраховано на просту структуру змісту і навіть у цьому випадку не буває без глюків. Тому тут не будемо її чіпати. З нею та іншими варіантами творення навігації можна ознайомитися на сайті <http://www.djvu-soft.narod.ru/>.

Розгляньмо творення навігації руками.

Для цього треба мати невелику утиліту **EmbeddedBookmarks-1.0**, що її створив Andrew Zhezherun. Вона вставляє у djvu-файл інформацію з спеціального html-файлу.

Цей файл має просту структуру і його можна створити вручну у будь-якому текстовому редакторі, наприклад, у **Notepad**. Його структуру можна зрозуміти з такого прикладу:

```
<body>
<li><a href>Рядок без посилання. В таких рядках можна подати
назву та автора</a> </li>
<ul>
<li><a href="#1">Посилання на сторінку 1. Зазвичай це обкладинка
книжки.</a></li>
<li><a href="#2">Посилання на сторінку 2</a></li>

<li><a href="#3">Частина I</a>
  <ul>
    <li><a href="#4">Розділ 1 на сторінці 4</a></li>
    <li><a href="#5">Розділ 2 на сторінці 5</a></li>
  </ul></li>
<li><a href="#6"> Частина II на сторінці 6</a>
  <ul>
    <li><a href="#6">Розділ 2</a>
      <ul>
        <li><a href="#7">Підрозділ на сторінці 7</a></li>
        <li><a href="#8">Підрозділ на сторінці 8</a></li>
      </ul></li>
    <li><a href="#9">Частина 3</a></li>
  </ul></li>
</ul>
</body>
</html>
```

Тут `` та `` — початок та кінець рядка.

`` та `` — початок та кінець переліку.

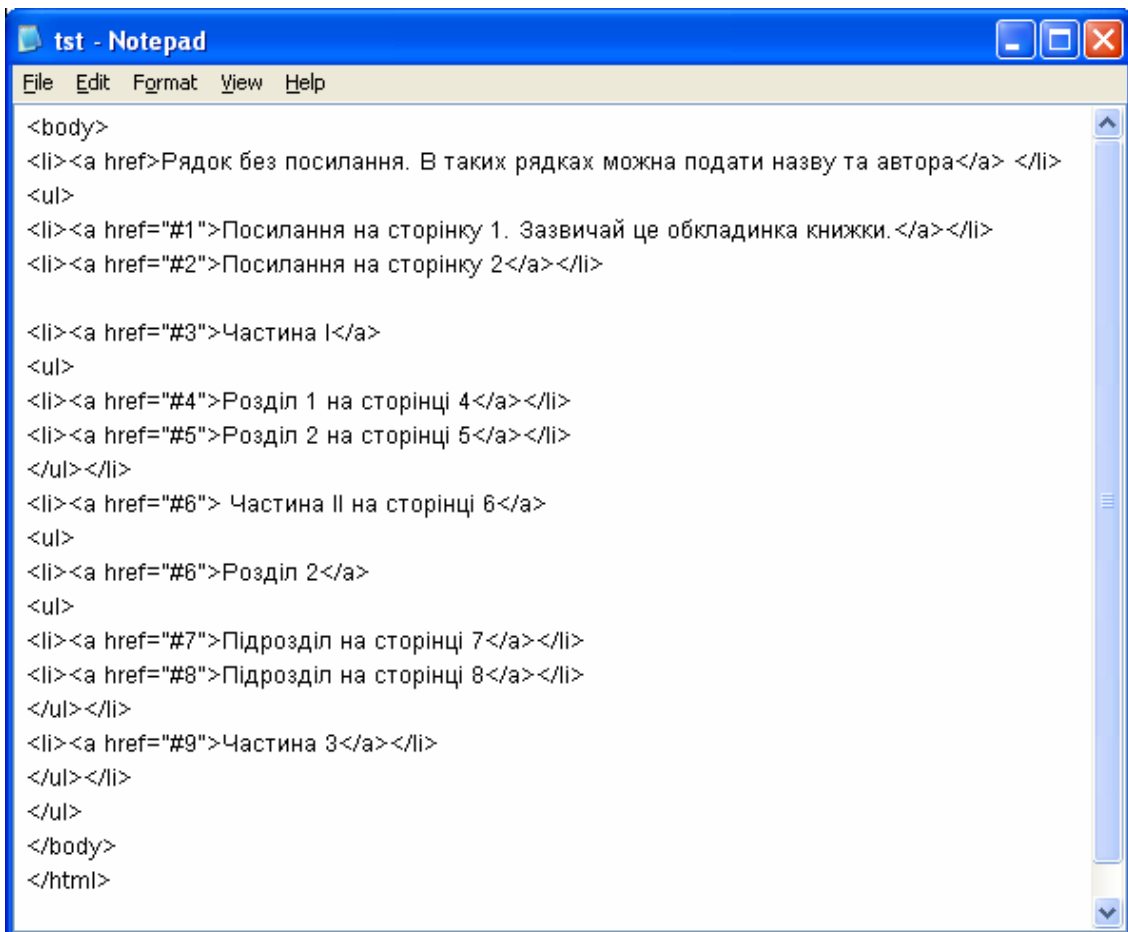
`` та `` — початок та кінець гіперпосилання на сторінку N. Між цими тегами стоїть текст, що його буде бачити користувач.

Заувага. Рядок без гіперпосилання не буде відображений, тому текст, що не має посилань на сторінку треба оформлювати як порожнє гіперпосилання (перший рядок прикладу):

```
<li><a href>ТЕКСТ</a> </li>.
```

Далі покрокова інструкція.

Набираємо (або копіюємо) цей текст у **Notepad**



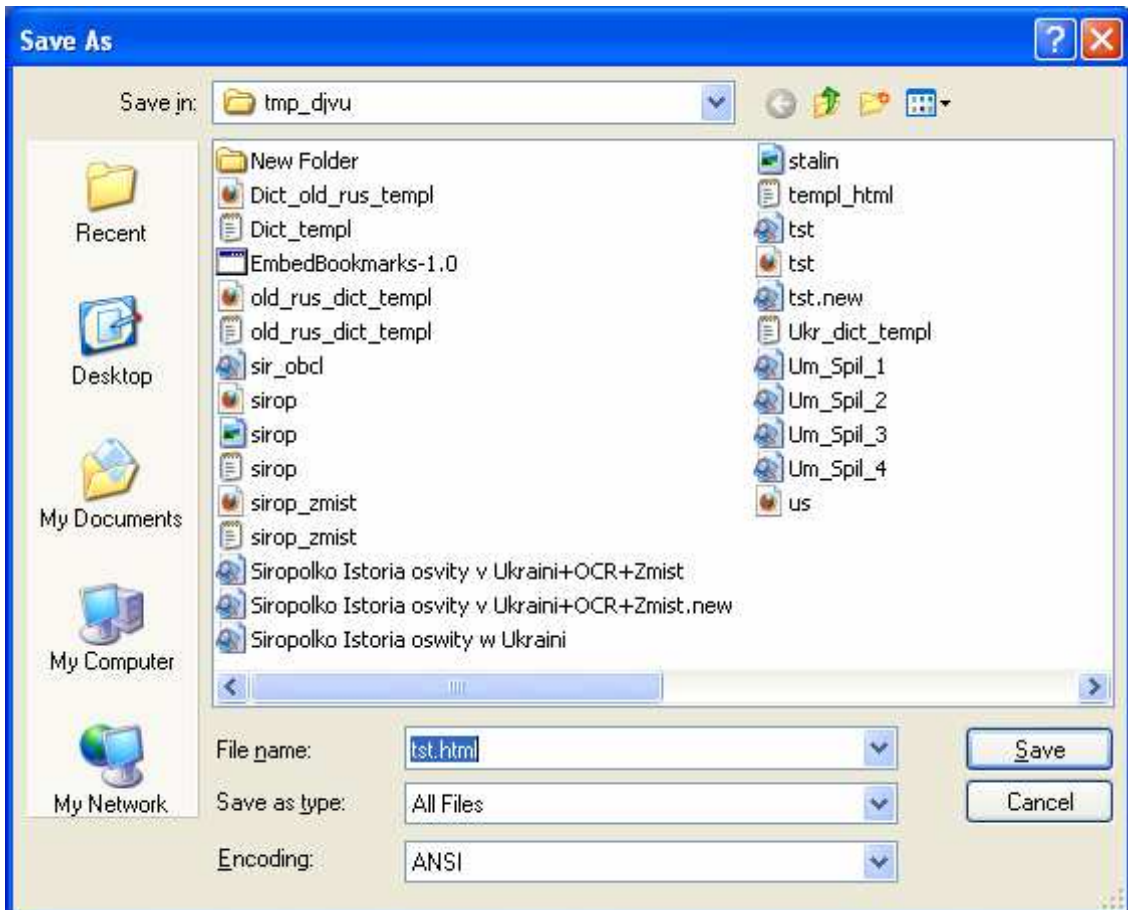
tst - Notepad

File Edit Format View Help

```
<body>
<li><a href>Рядок без посилання. В таких рядках можна подати назву та автора</a> </li>
<ul>
<li><a href="#1">Посилання на сторінку 1. Зазвичай це обкладинка книжки.</a></li>
<li><a href="#2">Посилання на сторінку 2</a></li>

<li><a href="#3">Частина I</a>
<ul>
<li><a href="#4">Розділ 1 на сторінці 4</a></li>
<li><a href="#5">Розділ 2 на сторінці 5</a></li>
</ul></li>
<li><a href="#6"> Частина II на сторінці 6</a>
<ul>
<li><a href="#6">Розділ 2</a>
<ul>
<li><a href="#7">Підрозділ на сторінці 7</a></li>
<li><a href="#8">Підрозділ на сторінці 8</a></li>
</ul></li>
<li><a href="#9">Частина 3</a></li>
</ul></li>
</ul>
</body>
</html>
```

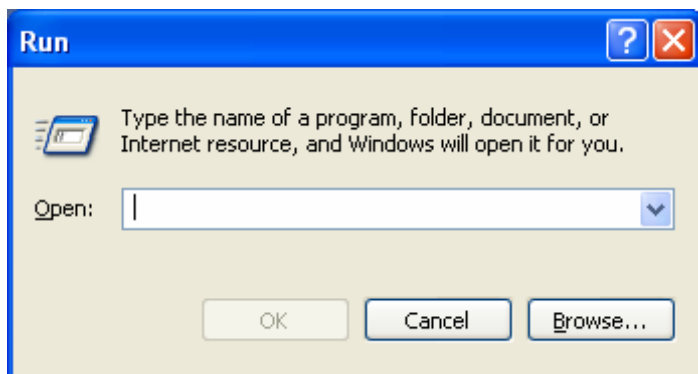
і зберігаємо його з розширенням **html**.



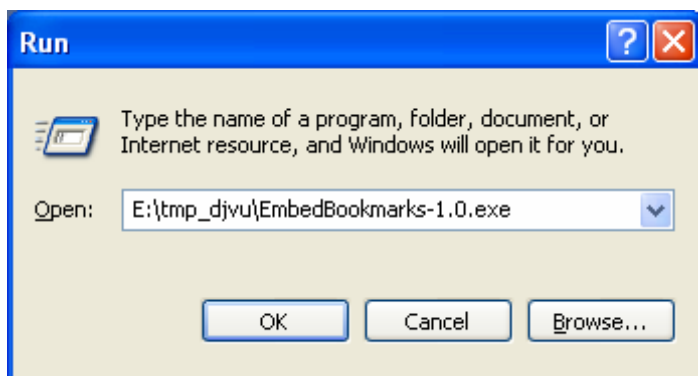
Зверніть увагу, що у віконці **Save as type** має стояти *All Files*, а в **Encoding** — *ANSI* або *Unicode*, якщо Ви користуєтеся юнікодними фонтами.

Утиліта **EmbeddedBookmarks-1.0** має жити в тій же течці, що і щойно записаний файл та djvu-файл, куди ми вставляємо навігацію.

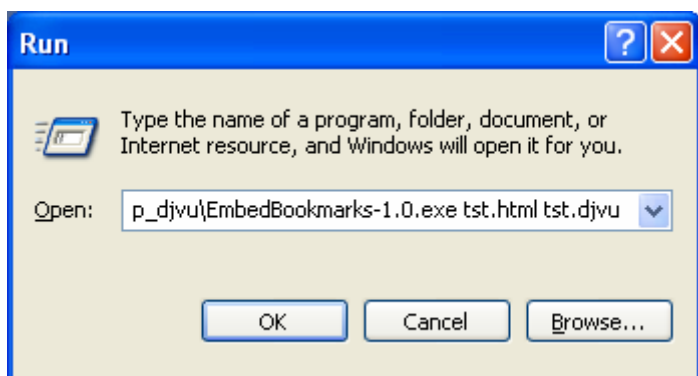
Тиснемо клавішу **Start** та обираємо **Run**. Бачимо діалог:



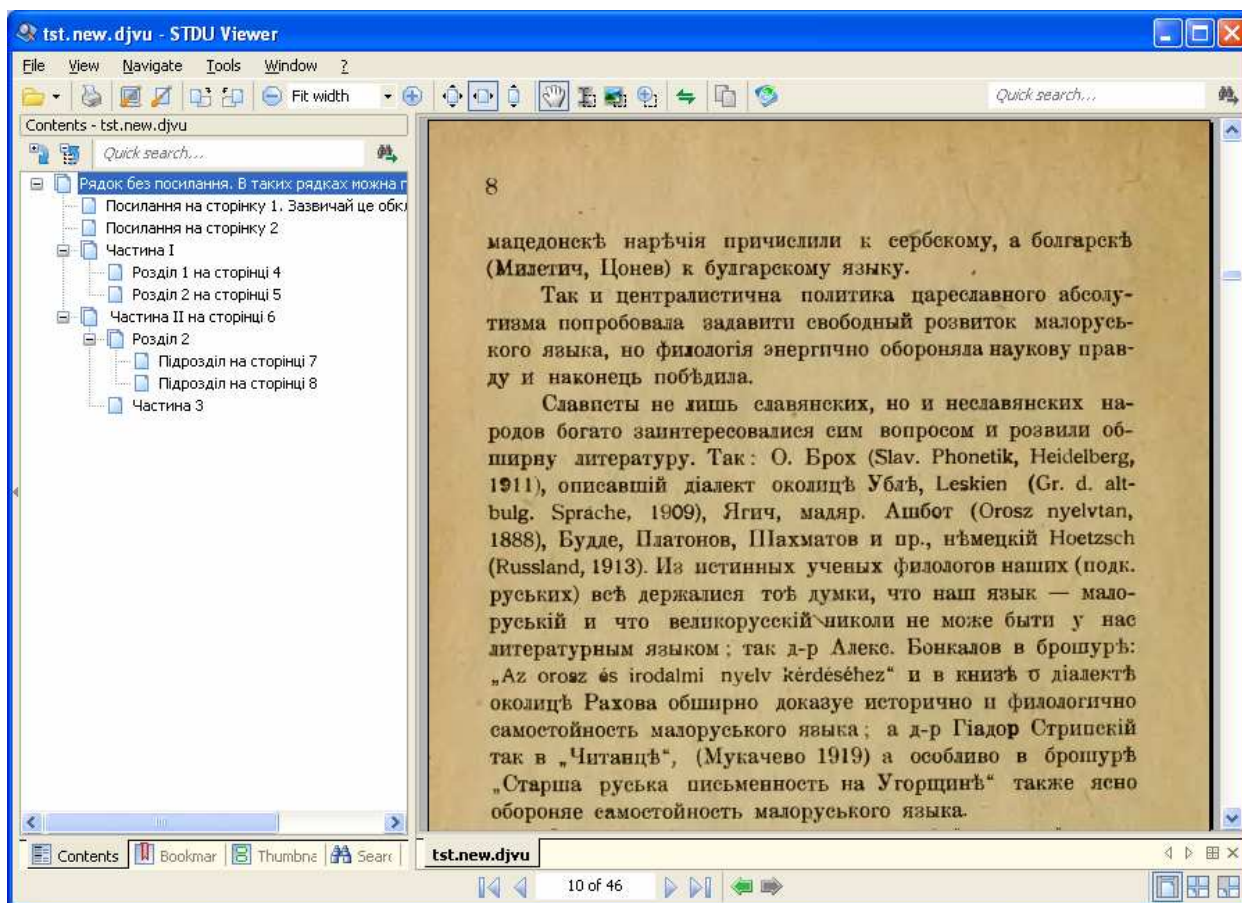
Натискаємо **Browse** і знаходимо нашу течку. Виділяємо **EmbeddedBookmarks-1.0**. В результаті отримуємо діалог.



В отриманому діалозі дописуємо, через пробіл, назву **html** та **djvu** файлів (звичайно, назви файлів можуть бути довільними):



Тиснемо **OK** і бачимо, що в течці з'являється новий файл, що має назву **FileName.new.djvu**. У даному випадку **tst.new.djvu**. Це файл з навігацією. Розкриваємо його і бачимо в лівій частині вікна очікуваний зміст.



От тепер справді все. Ми отримали djvu-книжку з текстовим шаром та навігацією.

6. Де взяти програми

Тут freeware означає, що програма є продуктом вільним, зокрема й від оплати.

IrfanView	www.irfanview.com	freeware
ScanKromsator	http://www.djvu-soft.narod.ru/soft/scan_kromsator_v5_92_full.rar	freeware
DjVu Small	http://www.djvu-soft.narod.ru/soft/djvu_small_v0_4_3.rar	freeware
ABBY FineReader	www.abbyy.com	trial
DjvuOCR 2.3	http://www.djvu-soft.narod.ru/soft/djvu_ocr_v2_3.rar	freeware
DjVu Hyperlinks Editor	http://www.djvu-soft.narod.ru/	freeware
EmbedBookmarks-1.0	http://sourceforge.net/projects/windjview/files/Bookmark%20Tool/1.0/EmbedBookmarks-1.0.exe/download	freeware

Для лінукоїдів існує рекомендація від **are** — набір програм **all2djvu**, що охоплюють увесь цикл створення djvu-книжок <http://www.djvu-soft.narod.ru/soft/all2djvu.htm>.

7. Тим, хто хоче йти далі

Дуже багато матеріалів щодо сканування, структури та оброблення djvu-файлів, програм та творення djvu-книжок є на сайті

<http://www.djvu-soft.narod.ru/>

А чи йти далі, вже справа кожного.

Успіхів у творенні добрих djvu-книжок!