Wikisource: come creare un file DjVu

Alex brollo

DjVu: cos’è

File multipagina di immagini
associate a uno strato testo ricercabile

- Aperto
- “Semplice”
- Leggero

Aperto, “semplice”, leggero

- Aperto: disponibile una libreria (DjvuLibre) per creazione, modifica, estrazione…. e un ottimo visualizzatore (DjView)

- “Semplice”: non proprio … ma più del PDF

- Leggero: implementa un formidabile meccanismo di compressione (opzionale)

Strato testo ricercabile

- Testo nascosto mappato in blocchi annidati

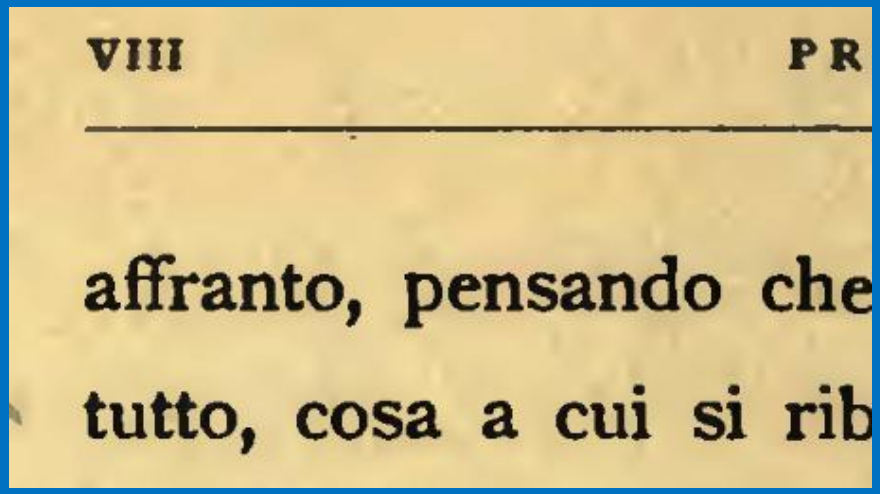
- Livelli di annidamento: PAGE, COLUMN, REGION, PARAGRAPH, LINE, WORD , CHAR

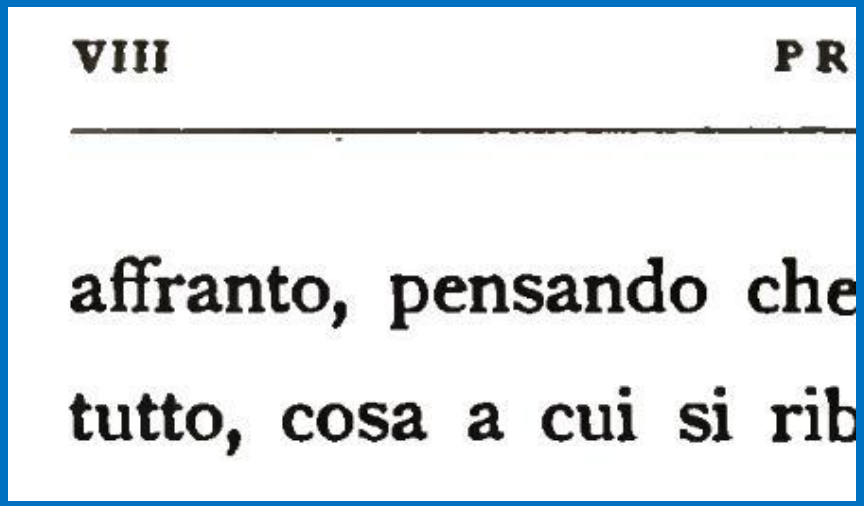
- Il testo mappato può essere estratto, modificato, e ricaricato con DjvuLibre.

Un buon djvu…

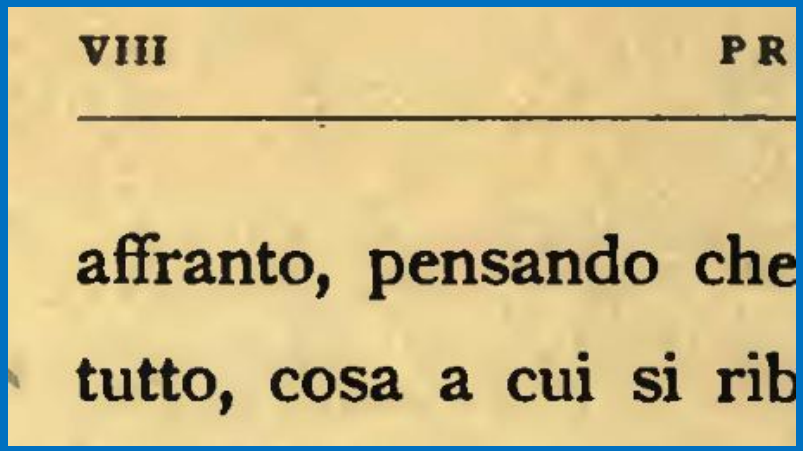
“montaggio” di buone immagini + buon testo mappato (ossia buon OCR)

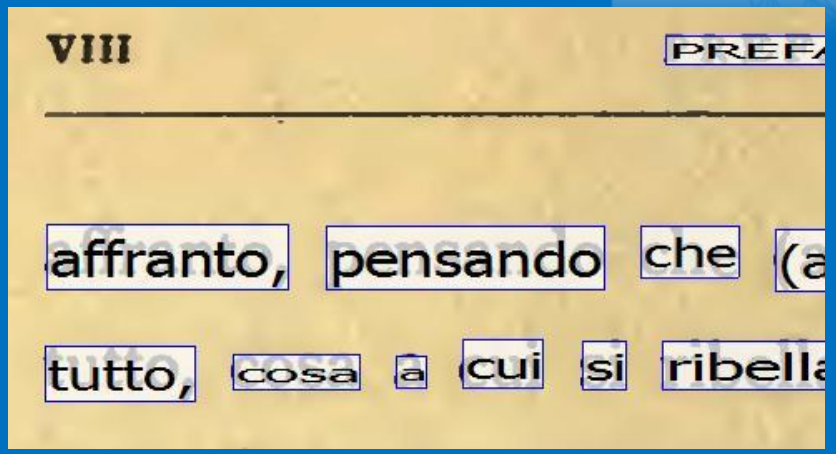
Struttura djvu: foreground-background

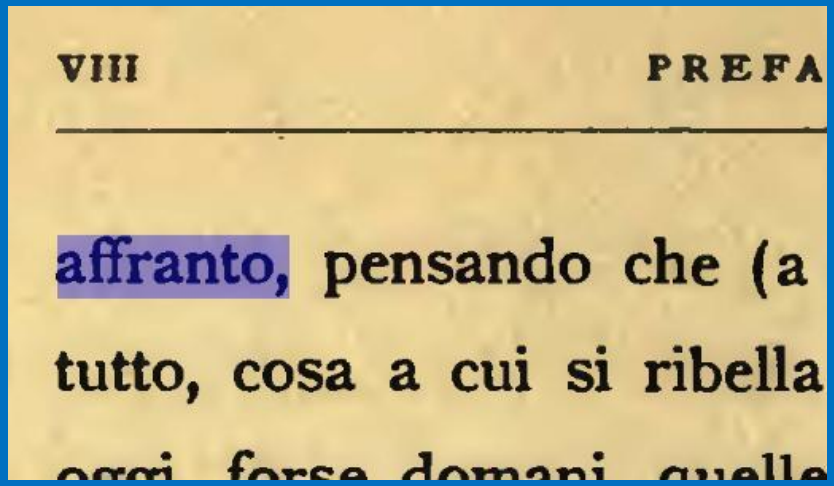


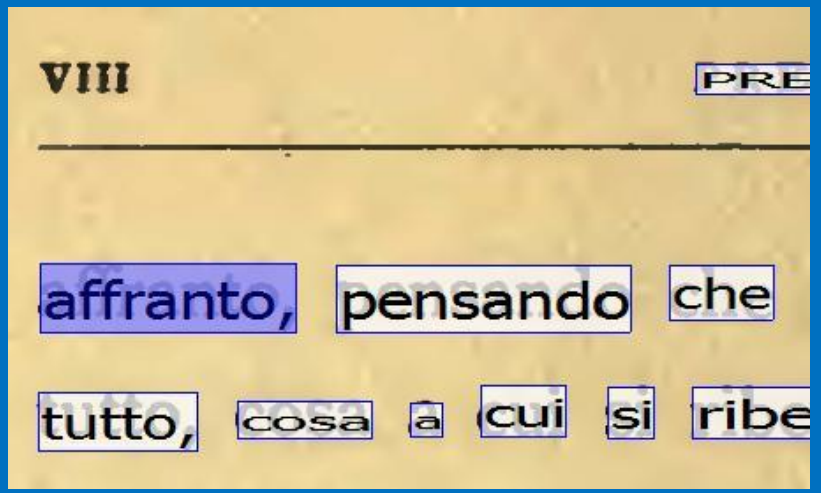


Struttura djvu: testo mappato ricercabile









Tutto in un colpo solo

Per costruire un buon djvu completo (o pochi):

Usare ABBYY FineReader ma:

- Commerciale
- Solo GUI (un testo alla volta)

Per creare molti djvu…

… avete bisogno di automatizzare, almeno parzialmente.

Fino a pochi mesi fa, la soluzione era: caricare su Internet Archive, che utilizza la versione *engine* di ABBYY FineReader.

Al momento Internet Archive non deriva più il formato djvu per i nuovi caricamenti.

I file di Internet Archive

Niente più djvu ma:

1. Pdf originale
2. File …._jp2.zip
3. File …._djvu.xml (testo OCR mappato)

- Pdf o _jp2.zip -> djvu “sole immagini”
- djvu “sole immagini” + _djvu.xml -> djvu con OCR

Costruire il djvu “sole immagini”

- Da pdf:
  - usare il servizio online any2djvu
  - usare pdf2djvu (solo Linux/unix o Win)

- Da immagini:
  - Usare DjvuToy
  - Usare il servizio online any2djvu
  - Usare le routine DjvuLibre

Montare il testo mappato

- Usare lo script python xml2dsed.py

(in itwikisource, Progetto:Bot/Programmi in Python per i bot)

Alternativa: IA Upload

Il tool IA Upload (https://tools.wmflabs.org/ia-upload/) può creare un djvu dai file IA:

- Costruisce un djvu immagini da …._jp2.zip di IA
- Monta nel djvu immagini l'OCR di …._djvu.xml di IA

Ma…. abbastanza spesso non ce la fa

Il “tool magico”

E’ disponibile un “tool magico”:

- Semplice
- Interattivo
- Interfaccia in linguaggio naturale

?

Risorse web

Internet Archive (IA)	https://archive.org
IA Upload	https://tools.wmflabs.org/ia-upload/
DjvuLibre	http://djvu.sourceforge.net/doc/index.html
Any2djvu	http://djvu.org/any2djvu/
xml2dsed.py	https://it.wikisource.org/wiki/Progetto:Bot/Programmi in Python per i bot/xml2dsed.py

Applicazioni offline

ABBYY FineReader

DjvuToy