Digitalizzare libri

Dal testo cartaceo al file searchable;
dal file searchable all’ipertesto.

Milano, 10 settembre 2019

Digitalizzare libri

Dal testo cartaceo al file searchable;
dal file searchable all’ipertesto.

Parte I: “File” searchable

Raccolta di immagini di pagine, in cui si può eseguire ricerche nel testo con evidenziazione delle parole risultanti.

Opere Complete di Goldoni, IV volume

Per molte dimostrazioni sarà usato il caso di “Opere complete di Carlo Goldoni”, volume IV, Municipio di Venezia, 1909.

La breve storia del volume.....

Internet Archive (archive.org)


OPERE COMPLETE

DI

CARLO GOLDONI

EDITE DAL MUNICIPIO DI VENEZIA

NEL II CENTENARIO DALLA NASCITA

Venezia

MDCCCCIX

Tre tipi di “file” seachable

Veri file:

- DjVu
- PDF

Applicazioni:

- visualizzatore di Internet Archive (jpg+xml)

La prima regola

Ottenere un buon file “searchable” è il primo passo per ottenere un buon ipertesto.

Corollario: ottenere e pubblicare sul web un buon file searchable non costituisce una “digitalizzazione completa” ma è già un ottimo risultato.

PDF

- Formato “open”, orientato alla stampa
- Estremamente potente e altrettanto complesso
- Non richiede software proprietario per la visualizzazione e la stampa (viene aperto anche dai principali browser web)
- Può essere utilizzato come semplice “contenitore di immagini”
- Può contenere anche immagini con sottostante testo mappato con possibilità di ricerca e evidenziazione di frammenti di testo (file “searchable”)

Esplorazione PDF

Sessione “live”:

- esplorazione di Opere complete di Goldoni - Vol. IV

DjVu

- Formato “open”, orientato alla visualizzazione di libri
- Struttura “semplice” e leggera
- Esistono plugin free per la sua visualizzazione con i principali browser e applicazioni free per la sua manipolazione (struttura, testo, immagini)
- Specificamente progettato per raccolte di immagini con sottostante testo mappato con possibilità di ricerca e evidenziazione di frammenti di testo (file “searchable”)
- Può essere anche utilizzato come semplice “contenitore di immagini”

Esplorazione DjVu

Sessione “live”:

- esplorazione di Opere complete di Carlo Goldoni

Visualizzatore IA

- Non è un file ma un’applicazione web
- Si basa sull’elaborazione contemporanea di una sequenza di immagini jpg e di un file xml con la rappresentazione mappata del testo
- Dispone di una gradevole animazione “flip” nello scorrimento delle immagini
- Permette la ricerca e l’evidenziazione di frammenti di testo
- Consente l’esportazione dell’intero libro e condivide i file sorgente!

Esplorazione visualizzatore IA

Sessione “live” via web:

- esplorazione di Opere complete di Goldoni - Vol. IV

Il percorso

Per ottenere un buon file searchable occorre seguire una buona procedura:

- buone immagini delle pagine (scansioni o fotografie);
- buona post-elaborazione delle immagini

Poi:

- ottimo OCR
- conversione delle immagini + OCR in un buon file searchable

oppure:

- caricamento su Internet Archive

L’attrezzatura.....

Programmi utili

Pdfsam	Divisione e fusione di pdf
Briss	Ritaglio rapido immagini pdf multipagina
XnView	Manipolazione immagini (anche pdf singoli)
pdftopng	Estrazione immagini (da pdf a png)
Scan Tailor	Normalizzazione di immagini da scansione (non pdf singoli)
DjvuLibre	Manipolazione di file DjVu
ABBYY Finereader	Processazione completa dalle immagini a vari tipi di file (anche file seachable)

Buone immagini

- riproduzione completa (copertina e tutte le pagine bianche!)
- adeguata risoluzione (almeno 300 dpi; non esagerare!)
- buon contrasto (caratteri scuri e fondo chiaro)
- limitazione al minimo degli artefatti
- salvataggio in file di alta qualità (meglio lossless: tiff, png)
- se il formato del libro lo permette, oppurtuna scansione a doppia facciata per dimezzare i tempi

Buona post-elaborazione

- verifica di completezza
- verifica dell’orientamento
- splitting delle pagine nelle scansioni a doppia pagina
- raddrizzamento (deskewing)
- ritaglio
- eventuale correzione di artefatti
- eventuale conversione in BN mediante Adaptive Thresholding
- verifica ed eventuale modifica dei nomi dei file immagine: numerazione zero-filled! MAI -1, -2, ...-10 ma -001, -002, ...-100

Scan Tailor (non pdf)

- verifica di completezza
- verifica dell’orientamento
- splitting delle pagine nelle scansioni a doppia pagina
- raddrizzamento (deskewing)
- ritaglio
- eventuale correzione di artefatti (alcuni)
- eventuale conversione in BN mediante Adaptive Thresholding

Briss (pdf multipagina!)

- verifica di completezza
- splitting delle pagine nelle scansioni a doppia pagina
- verifica dell’orientamento
- raddrizzamento (deskewing)
- ritaglio
- eventuale correzione di artefatti
- eventuale conversione in BN mediante Adaptive Thresholding

ABBYY FineReader

- verifica di completezza
- splitting delle pagine nelle scansioni a doppia pagina
- verifica dell’orientamento
- raddrizzamento (deskewing)
- ritaglio
- eventuale correzione di artefatti
- esecuzione OCR di altissima qualità
- esportazione in pdf o djvu searchable
- commerciale, salvataggio file sorgente in formato proprietario

In pratica

- Il fattore limitante della scansione e della postelaborazione è il tempo-uomo .
- E’ opportuno un investimento iniziale di tempo nello studio delle opzioni dello scanner e nell’autoaddestramento sui programmi per la postelaborazione.
- La procedura più semplice e rapida è il salvataggio diretto delle scansioni a facciata singola o doppia su pdf multipagina seguito da verifica della completezza, splitting eventuale e ritaglio con briss, verifica del risultato.
- Se il risultato non è soddisfacente.....

Post-elaborazione completa

- Se è sufficiente la sequenza orientamento, splitting, raddrizzamento e ritaglio: estrazione delle immagini in formato grafico lossless (tiff, png) e elaborazione con Scan Tailor

- Se è opportuna una elaborazione digitale complessa delle immagini (contrasto, tonalità, cancellazioni....): conversione pdf multipagina in file grafici standard (tiff, png) e elaborazione immagini con programmi capaci di elaborazioni batch (es. XnView)

Infine: caricare su IA

- La procedura di caricamento su Internet Archive è gratuita e piuttosto semplice (basta la registrazione).
- L’uploader rimane “amministratore di ciò che ha caricato”: mantiene la facoltà di modifica, sostituzione dei file, rielaborazione sui propri item.
- I metadati obbligatori sono pochissimi; quelli opzionali sono moltissimi.
- Il caricamento di pdf multipagina è molto più semplice del caricamento alternativo (archivio zip della sequenza delle immagini).

Dopo caricato su IA...

- Se il caricamento su IA si completa con successo, parte immediatamente un processo di derivazione con analisi delle immagini e esecuzione di un eccellente OCR (mediante ABBYY FineReader).
- Al termine della derivazione (qualche ora) il libro può essere esaminato con il visualizzatore.
- Possono (e dovrebbero!) essere scaricati tutti i file originali e tutti i file derivati, compresi il risultato di massimo dettaglio dell'OCR in formato non proprietario (xml) e la raccolta delle immagini di alta qualità.
- L'output di IA è utilizzabile (direttamente o previa elaborazione) come input per ottenere un ipertesto (Distributed Proofreader, che alimenta il progetto Gutenberg; Wikisource; altri progetti).

Conclusione parte I

Ottenere la trasformazione di un libro cartaceo in immagini searchable è un ottimo risultato. Richiede un esiguo impegno economico, ma un discreto investimento in tempo-uomo.

Per vari motivi, la pubblicazione sul web utilizzando la piattaforma gratuita offerta da Internet Archive è fortemente raccomandabile.

Parte II: Ipertesti

Tralascio la definizione di ipertesto. Il più comune formato di ipertesto è l'HTML, e la trasformazione di un testo cartaceo in ipertesto in sostanza è la sua rappresentazione HTML che consente:

- qualsiasi ricerca sul testo;
- analisi del testo;
- collegamenti (links) e inserimento di immagini;
- esportazione in una varietà di formati.

Il ruolo di wikisource

Wikisource:

- raccoglie testi liberi da copyright
- li rielabora in ipertesti
- li trasforma in nodi della rete dei progetti wiki

My pain: esistono molte wikisource, una per lingua, completamente indipendenti e oggetto di evoluzione divergente.

La procedura *proofread* di wikisource

La procedura proofread di wikisource utilizza preferenzialmente file *searchable* (DjVu, PDF) come input.

Il primo passo della trasformazione in ipertesto è la correzione del testo OCR usualmente incorporato nei file *searchable*.

Al momento:

- una notevole percentuale di file *searchable* proviene da Internet Archive;
- viene utilizzata una minima parte delle informazioni prodotte dalla procedura OCR.

Passo 1: correzione dell’OCR

L’OCR ha raggiunto notevole affidabilità ma siamo molto lontani dalla perfezione:

- la sua accuratezza si riduce progressivamente con i testi più vecchi;
- il fatto di basarsi su vocabolari della lingua moderna introduce errori nel riconoscimento di varianti ortografiche desuete;
- gli errori si concentrano nelle parti più significative del testo: citazioni, nomi di autori, abbreviazioni.
- Ogni ricerca/elaborazione testuale basata solo sull’OCR è incompleta e quindi inaffidabile.

Si fa presto a dire OCR

- OPERE

- <WORD coords="826,886,1626,694,883">OPERE</WORD>

- <charParams l="826" t="694" r="1006" b="882" wordFirst="1" wordLeftMost="1" wordFromDictionary="1" wordNormal="1" wordNumeric="0" wordIdentifier="0" charConfidence="92" serifProbability="255" wordPenalty="0" meanStrokeWidth="279">O</charParams>

  <charParams l="1014" t="694" r="1158" b="882" wordFromDictionary="1" wordNormal="1" wordNumeric="0" wordIdentifier="0" charConfidence="44" serifProbability="44" wordPenalty="0" meanStrokeWidth="279">P</charParams>

  <charParams l="1162" t="698" r="1306" b="882" wordFromDictionary="1" wordNormal="1" wordNumeric="0" wordIdentifier="0" charConfidence="35" serifProbability="100" wordPenalty="0" meanStrokeWidth="279">E</charParams>

  <charParams l="1310" t="698" r="1478" b="886" wordFromDictionary="1" wordNormal="1" wordNumeric="0" wordIdentifier="0" charConfidence="65" serifProbability="64" wordPenalty="0" meanStrokeWidth="279">R</charParams>

  <charParams l="1482" t="698" r="1626" b="886" wordFromDictionary="1" wordNormal="1" wordNumeric="0" wordIdentifier="0" charConfidence="29" serifProbability="100" wordPenalty="0" meanStrokeWidth="279">E</charParams>

Passo 2: formattazione

- Oltre ad una esatta riproduzione del testo (caratteri e parole), sono necessari, per la sua fruizione:
  - una adeguata struttura (divisione in paragrafi, in versi, spaziature verticali...)
  - accorgimenti di stile (grandezza dei caratteri, grassetto, corsivo, apice/pedice...)
  - inserimento delle immagini

    e infine
  - inserimento delle annotazioni e dei link interni ed esterni

Uno sguardo al risultato

Come nel caso dei file “searchable”, esaminiamo il risultato finale prima di illustrare il percorso per la sua realizzazione

Rappresentazione digitalizzata dell’intero volume:

- Opere complete di Carlo Goldoni - Vol. I
- Opere complete di Carlo Goldoni - Vol. IV

Rappresentazione digitalizzata di una singola commedia:

- Terenzio

Il percorso

Si parte da un file searchable (djvu o pdf) pubblicato (Internet Archive!) o autoprodotto.

- Il file viene caricato su Commons

  Opere complete di Goldoni, vol. IV

- In wikisource si crea la pagina base Indice collegata

  Opere complete di Goldoni, vol. IV

- In wikisource, pagina per pagina, si eseguono i due passaggi di correzione del testo e formattazione, e validazione finale
- segue la transclusione in un testo esportabile, organizzato in capitoli.

Sessione web: la pagina Indice

Esame approfondito di Opere complete di Goldoni, vol. IV

- pagina Indice

Sessione web: le pagine Pagina

Esame approfondito di pagine dell’opera, level 1

- pagina 215 del volume IV

Esame di una pagina level 3 (riletta e formattata)

- Pagina 11 del Vol. VII, da Il Moliere

Sessione web: la pagina testo

Esame di “Il Moliere”

- Il Moliere

Esportazione in pdf (attualmente malfunzionante!)

- Il Moliere.pdf

I problemi nella trascrizione

- Tempo - uomo! Eccellenza: 120 sec./pagina = 30 pagine/h
- I tempi per pagine complesse sono molto superiori
- Necessario un intenso auto-addestramento per esplorare e utilizzare i numerosi tools disponibili
- Scarsa documentazione, rapida evoluzione del software di base
- Scarsa condivisione di convenzioni e soluzioni fra i vari progetti
- Interfaccia adattata e non completamente dedicata

Il piacere della trascrizione

- opportunità di lettura molto attenta di testi interessanti
- opportunità di affrontare problemi informatici non banali
- opportunità di frequentare una comunità particolarmente amichevole e collaborativa

Statistiche di attività

- Le statistiche di attività ProofreadPage dispongono di un tool dedicato:

  https://tools.wmflabs.org/phetools/statistics.php

Conclusione

La trasformazione da file searchable a ipertesto è complessa e richiede molto tempo-uomo.

- E’ irrealistico immaginare che wikisource possa completare la digitalizzazione di tutti i testi trasformati in file *searchable*
- E’ irrealistico immaginare che la proposta di digitalizzare un libro specifico possa essere rapidamente accolta ed evasa, senza fornire un’adeguato “tempo-uomo” di supporto.
- La digitalizzazione fino al file searchable è comunque un obiettivo importante e dignitoso, che andrebbe energicamente e diffusamente perseguito.