OCR – RICONOSCIMENTO OTTICO DEI CARATTERI PER MATERIALI A STAMPA

Per accedere ai contenuti testuali delle immagini prodotte dalle scansioni digitali dei volumi e documenti a stampa è necessario l'OCR.

Allo stato attuale sono possibili tre opzioni:

a) l'elaborazione di più file PDF che riconsegnano l'informazione della presenza della parola a livello di pagina;
b) l'elaborazione di più file TXT che riconsegnano l'informazione della presenza della parola a livello di pagina;
c) l'elaborazione di unico file XML elaborato sullo Schema ALTO[1] che raggiunge attraverso le coordinate la parola cercata nella sua posizione su tutto il documento.

Procedura per opzione A

Sul TIFF sarà generato un file in formato PDF con OCR enbedded, per ogni file immagine dell'intero documento. Nella cartella digitale sarà collocata la cartella contenente i file pdf prodotti tramite il software per il riconoscimento ottico, vedi anche doc. File System.

Del file PDF generato, Internet Culturale gestisce il livello (layer) "testo", infatti la parte immagine del PDF non viene fruita tramite l'utilizzo del viewer del portale.

Quindi, il settaggio del layer "immagine" deve essere impostato in modo che il file non sia eccessivamente "pesante". Utilizzare un formato immagine compresso, una risoluzione a 72 dpi, una profondità di colore 8 bit per versioni in scala di grigio o 24 bit per immagini a colore (RGB).

In ogni caso, il"peso" del singolo file PDF non dovrà superare il valore di 1 mb.

Nel MAG, viene creata la sez. OCR con i file pdf che riporteranno la stessa nomenclatura, lo stesso sequence_number e nome del file immagine corrispondente nella sez. IMG e lo usage 3 che è quello dedicato alla copia per il web, per Internet culturale. Non creare il file pdf per la scansione dedicata alla color chart.

Procedura per opzione B

[1] http://www.loc.gov/standards/alto/news.php

Sul TIFF sarà generato un file in formato TXT per ogni file immagine dell'intero documento. Nella cartella digitale sarà collocata la cartella contenente i file TXT prodotti tramite il software per il riconoscimento ottico.

Nel MAG, viene creata la sez. OCR con i file TXT che riporteranno la stessa nomenclatura, lo stesso sequence_number e nome del file immagine corrispondente nella sez. IMG e lo usage 3 che è quello dedicato alla copia per il web, per Internet culturale. Non creare il file TXT per la scansione dedicata alla color chart.

La scelta tra le opzioni A e B deve essere effettuata sui materiali oggetto della digitalizzazione dal responsabile del progetto su campioni di prova forniti dal digitalizzatore per verificare la qualità e individuare il risultato migliore tra i due formati.

Procedura per opzione C

la produzione di un unico file xml elaborato in base allo schema ALTO di codifica, sia in grado di restituire le coordinate di ogni singola parola direttamente sul file di ciascuna immagine di usage 3 (sez. IMG per WEB).
Nella cartella digitale sarà collocata la cartella contenente l'unico file xml contenente il testo prodotto dal riconoscimento ottico codificato in base allo standard ALTO e le coordinate, generato dal TIFF

Nel MAG viene creata la sez. OCR con l'unico file descritto che rimanda nel suo path alla cartella digitale. Non considerare la scansione dedicata alla color chart.

Anche la sez. DOC è dedicata alla descrizione di file di testo e raccoglie file di testo born digital oppure testi che derivano da OCR ma che sono stati sottoposti a controllo editoriale manuale.