Specifiche rispetto alle attività di digitalizzazione e post-produzione progetto di digitalizzazione PNRR M1C3 1.1.5 -Quota Regionale

Procedura digitalizzazione Carta e Foto

- **Caratteristiche File Master TIFF – compressione LZW.** Al fine di ottimizzare il peso dei file prodotti e facilitare le operazioni di storage e di ingestion, occorre applicare a tutti i file Master TIFF prodotti **l'algoritmo di compressione lossless LZW** (Lempel-Ziv-Welch).

- **Mosaicatura.** Per quei beni con misure che implicherebbero una ripresa da distanza non raggiungibile all'interno degli spazi messi a disposizione dagli enti, è ammessa la restituzione dell'intero disegno attraverso la mosaicatura (stitching) di più acquisizione fotografiche eseguite secondo le specifiche tecniche indicate nelle condizioni generali.

- **Specifiche rispetto all'attività di lettura OCR.** Il formato richiesto è hOCR, standard aperto per la rappresentazione dei dati OCR (Optical Character Recognition) associati alle immagini. Basato su XML, hOCR utilizza la struttura di HTML o XHTML per codificare le informazioni estratte, combinando flessibilità e facilità d'uso con una solida base tecnologica ampiamente supportata.

  È fondamentale che l'output hOCR contenga le seguenti informazioni principali, da considerarsi **obbligatorie ai fini del progetto**:

  1. Il titolo principale (main title) della pagina, se presente;
  2. Una o più aree di testo, corrispondenti a eventuali colonne presenti nella pagina, rispettando l'ordine di lettura;
  3. Per ogni area, una struttura organizzata in linee di testo, ciascuna contenente le parole riconosciute. L'ordine di lettura deve essere mantenuto sia per le linee che per le parole all'interno di esse.

  Inoltre, è possibile **facoltativamente** inserire ulteriori metadati, di seguito elencati:

  1. Il numero della pagina corrente rispetto al totale delle pagine del documento;
  2. La suddivisione del testo in paragrafi all'interno di ogni area, utilizzando un'apposita classe HTML indicata nel prosieguo del presente documento;
  3. L'identificazione di eventuali titoli secondari;
  4. Il livello di confidenza della predizione OCR per ogni singola parola.

  Questa struttura assicura una rappresentazione accurata e dettagliata del contenuto testuale estratto dall'immagine, facilitando successive elaborazioni e analisi del documento digitalizzato.

Specifiche del profilo

Il formato hOCR indicato utilizza una struttura basata su HTML con specifici elementi e attributi per codificare le informazioni di OCR. Un esempio grafico dell'output ottenuto dall'OCR di una pagina di quotidiano è il seguente:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN"
  "http://www.w3.org/TR/html4/loose.dtd">
<html>
<body>
  <div class='ocr_page' id='page_1' title='image "example.png"; bbox 0 0 2550 3300; ppageno 0'>
    <div class='ocr_carea' id='block_1_1' title="bbox 200 400 2350 800">
      <p class='ocr_title' id='title_1' title="bbox 200 400 2350 800">Main Title</p>
    </div>
    <div class='ocr_carea' id='block_1_2' title="bbox 200 900 1200 3200">
      <p class='ocr_par' id='par_1' lang='eng' title="bbox 200 900 1200 3200">
        <span class='ocr_line' id='line_1' title="bbox 200 900 1200 950">
          <span class='ocrx_word' id='word_1' title="bbox 200 900 300 950; x_wconf 29">.....</span>
          <span class='ocrx_word' id='word_2' title="bbox 310 900 410 950; x_wconf 86">.....</span>
        </span>
        <span class='ocr_line' id='line_2' title="bbox 200 960 1200 1010">
          <span class='ocrx_word' id='word_3' title="bbox 200 960 300 1010; x_wconf 97">.....</span>
          <span class='ocrx_word' id='word_4' title="bbox 310 960 410 1010; x_wconf 99">.....</span>
        </span>
      </p>
    </div>
    <div class='ocr_carea' id='block_1_3' title="bbox 1300 900 2350 3200">
      <p class='ocr_par' id='par_2' lang='eng' title="bbox 1300 900 2350 3200">
        <span class='ocr_line' id='line_3' title="bbox 1300 900 2350 950">
          <span class='ocrx_word' id='word_5' title="bbox 1300 900 1400 950; x_wconf 78">.....</span>
          <span class='ocrx_word' id='word_6' title="bbox 1410 900 1510 950; x_wconf 68">.....</span>
        </span>
        <span class='ocr_line' id='line_4' title="bbox 1300 960 2350 1010">
          <span class='ocrx_word' id='word_7' title="bbox 1300 960 1400 1010; x_wconf 93">.....</span>
          <span class='ocrx_word' id='word_8' title="bbox 1410 960 1510 1010; x_wconf 26">.....</span>
        </span>
      </p>
    </div>
  </div>
</body>
</html>
```

Di seguito, invece, è fornita una panoramica più dettagliata delle caratteristiche in esso contenute:

Struttura della pagina (ocr_page):

Ogni pagina del documento è rappresentata da un elemento div con classe ocr_page.

L'elemento ocr_page ha un attributo id univoco (ad esempio, page_1).

Contiene un attributo title che specifica il nome dell'immagine, le dimensioni della pagina (bbox) e facoltativamente il numero di pagina (ppageno).

Blocchi di area (ocr_carea):

Ogni colonna di testo o blocco principale è rappresentato da un element div con classe ocr_carea.

Ogni ocr_carea ha un attributo id univoco (ad esempio, block_1_1, ...).

L'elemento ocr_carea ha un attributo title che specifica la posizione (bbox) del blocco.

Paragrafi (ocr_par): (facoltativamente)

Ogni paragrafo all'interno di un blocco di area è rappresentato da un elemento p con classe ocr_par.

Ogni ocr_par ha un attributo id univoco (ad esempio, par_1, ...).

L'elemento ocr_par ha un attributo title che specifica la posizione (bbox) del paragrafo.
Può includere un attributo lang che specifica la lingua del testo.

Linee di testo (ocr_line):

Ogni linea di testo all'interno di un paragrafo è rappresentata da un elemento span con classe ocr_line.

Ogni ocr_line ha un attributo id univoco (ad esempio, line_1, ...).

L'elemento ocr_line ha un attributo title che specifica la posizione (bbox) della linea.

Parole (ocrx_word):

Ogni parola all'interno di una linea di testo è rappresentata da un elemento span con classe ocrx_word.

Ogni ocrx_word ha un attributo id univoco (ad esempio, word_1, ...,).

L'elemento ocrx_word ha un attributo title che specifica la posizione (bbox) della parola e facoltativamente la confidenza di predizione della parola (x_wconf).

All'interno di ogni marcatore, infine, sarà presente la parola riconosciuta.

Di seguito alcune caratteristiche generali costantemente presenti in ogni hOCR:

ID unici e crescenti:

Ogni elemento (ocr_page, ocr_carea, ocr_title, ocr_par, ocr_line, ocrx_word) ha un attributo id univoco e crescente per identificare in modo chiaro ogni parte del documento.

Posizioni specifiche (bbox):

Ogni elemento ha un attributo title che specifica le coordinate di bounding box (bbox) che delimitano la posizione dell'elemento all'interno dell'immagine del documento.

Organizzazione gerarchica:

La struttura gerarchica va dalla pagina (ocr_page), ai blocchi di area (ocr_carea), ai paragrafi (ocr_par), alle linee di testo (ocr_line) e infine alle parole (ocrx_word).

Allegati

Si allega al presente documento un esempio di file hOCR compilato secondo le specifiche indicate nel presente documento e la relativa immagine.

Restano valide tutte le altre indicazioni contenute nelle Condizioni Generali.