

PUBBLICAZIONE DI OGGETTI DIGITALI SULLA DIGITAL LIBRARY UNITO

INDICE

Preparazione del pacchetto di files per la DL Unito.....	2
Cosa bisogna ottenere.....	2
Google Cloud Storage e software “Cyberduck”.....	3
Installazione e configurazione del programma.....	4
UPLOAD DEI FILE TIFF su Google Cloud Storage.....	5
File COMMAND.cmd.....	8
Creazione dei file (vecchio metodo: senza usare Google Cloud Storage).....	9
File TIFF.....	9
File JPEG (sia a 150 che a 300 dpi):.....	10
File TXT.....	11
File PDF/A.....	14

Preparazione del pacchetto di files per la DL Unito

Cosa bisogna ottenere

Per poter procedere con la pubblicazione sulla Digital Library di Unito è necessario creare sul proprio PC locale un pacchetto di file (c.d. “di ingestion”) così composto:

- i file TIFF creati dallo scanner (ad alta risoluzione)
- un export dei file TIFF in JPEG (a 300 dpi)
- un export dei file TIFF in JPEG (a 150 dpi)
- i file di testo .txt generati dal processo di riconoscimento ottico dei caratteri
- un file PDF generato dai file TIFF

Tutti i file di cui sopra saranno organizzati in una **struttura di cartelle e sottocartelle come la seguente** (l'esempio si riferisce all'oggetto digitale con collocazione FP 0005, che ha BID=IEI0095335 e numero di inventario IFP 2862):

FP0005_IEI0095335_IFP2862		(cartella radice del volume nominata come collocazione_bid_inventario dell'oggetto digitale)
	tiff	(sottocartella con le immagini in formato .tiff provenienti dallo scanner nominati con numeri consecutivi (*) come indicato sopra)
	jpeg150	(sottocartella con le immagini in formato .jpg a 150 dpi nominati con numeri consecutivi come indicato sopra)
	jpeg300	(sottocartella con le immagini in formato .jpg a 300 dpi nominati con numeri consecutivi come indicato sopra)
	txt	(sottocartella con i files in formato .txt dell'OCR nominati con numeri consecutivi come indicato sopra)

	pdf	(sottocartella che contiene il file pdf di tutte le pagine del volume)
	IEI0095335.pdf	(file in formato PDF avente come nome il solo BID de volume e che contiene tutte le pagine del volume (**))

(*) - prestare attenzione che la cartella abbia nome “tiff” (con due f) e non ad esempio “tif”.

Prestare inoltre attenzione alla numerazione dei file che devono essere sequenziali da 0001 a nnnn senza salti. I salti possono essere presenti se quando si è svolta la scansione si sono cancellate immagini per rifarle in quanto venute male. Il processo di nomenclatura dei file da parte del software dello scanner nonostante i test effettuati non rinomina le pagine in modo corretto spesso mettendole al fondo.

()** - Il file pdf deve essere un unico file con tutte le pagine del volume e non i singoli file pdf delle pagine.

Per creare il pacchetto di cartelle e file come indicato sopra è disponibile una procedura automatizzata sulla piattaforma Google Cloud Storage. Vedere qui di seguito.

Google Cloud Storage e software “Cyberduck”

Google Cloud Storage è un uno spazio server di Google che **consente di velocizzare la creazione del pacchetto di Ingestion** per la Digital Library.

Caricando i soli file TIFF creati con lo scanner su Google Cloud Storage con pochi semplici passaggi esso è dotato di appositi script che automaticamente creeranno:

- i jpeg a 150 ppi
- i jpeg a 300 ppi
- i txt del riconoscimento ottico dei caratteri
- il pdf delle singole pagine partendo dai TIFF originali.

Google Cloud Storage può essere utilizzato anche per archiviare le digitalizzazioni (al posto di un hard disk esterno).

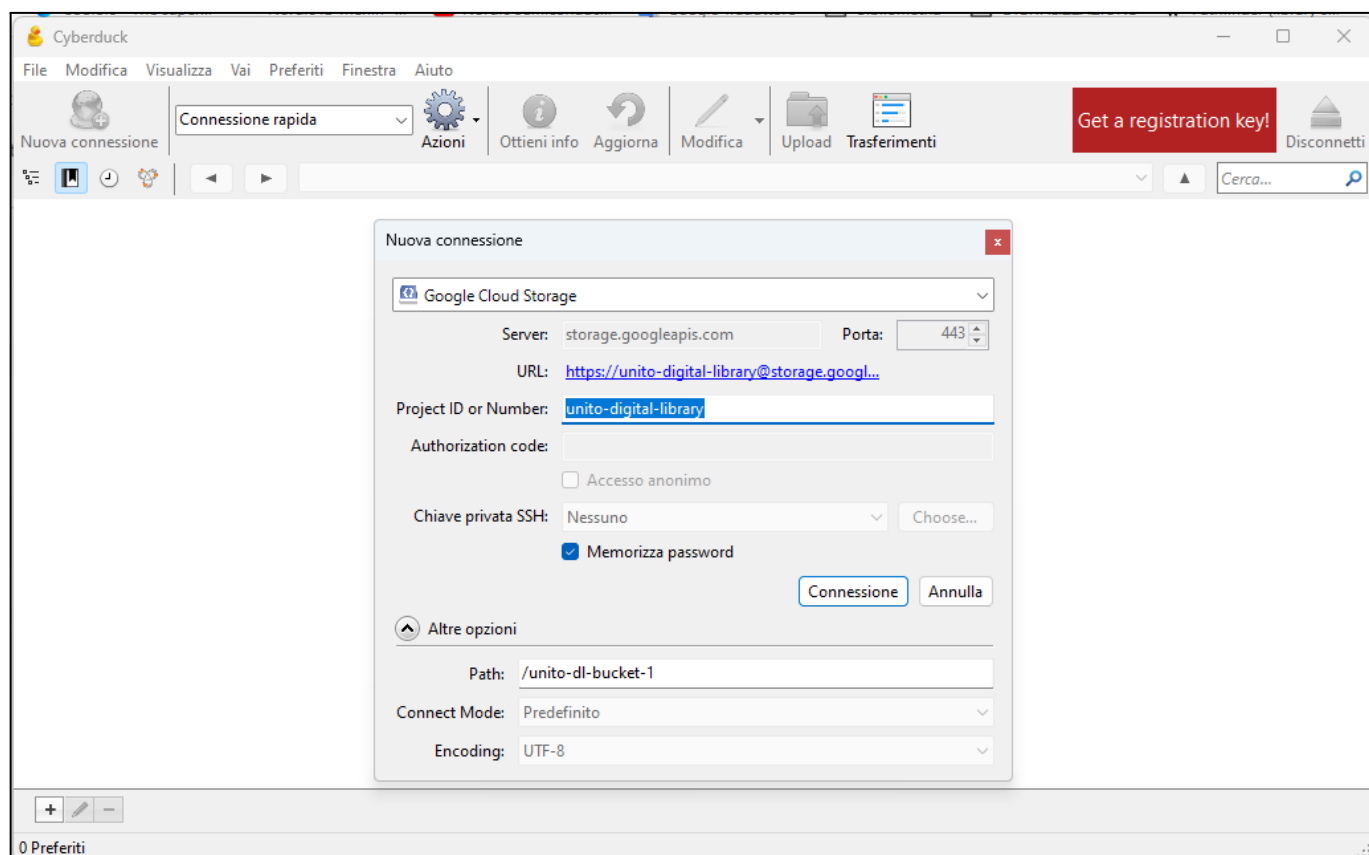
Installazione e configurazione del programma

Per usare Google Cloud Storage è necessario [scaricare ed installare il programma Cyberduck](#) (è sufficiente la versione gratuita).

Una volta scaricato ed installato (come qualunque altro programma) va configurato come segue:

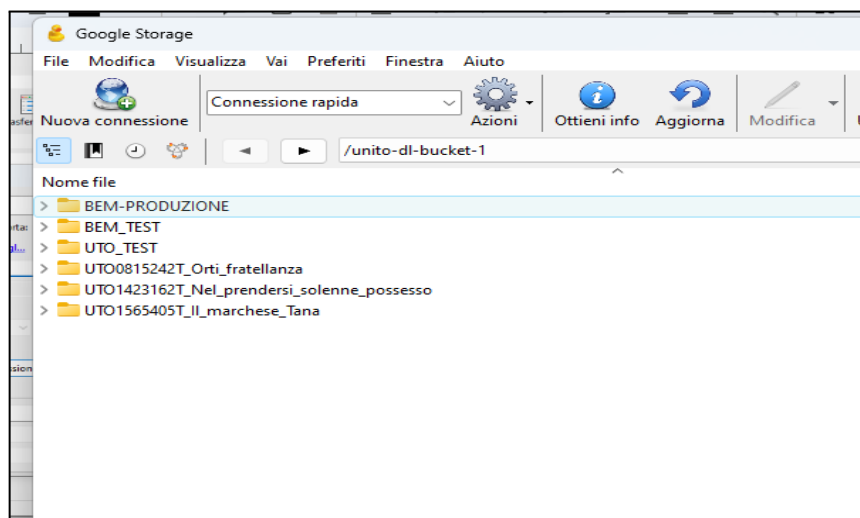
1. Aprire il programma e cliccare su “Nuova connessione”. Si apre una finestra dove impostare i **parametri che vengono indicati dai gestori del servizio**. Questi parametri sono diversi per ogni polo e **per la BEM** sono (vedere figura indicata sotto):
 - a. **options (path):** /unito-dl-bucket-1
 - b. **project id:** unito-digital-library
2. Cliccare su “Connessione” e seguire le istruzioni di autenticazione con la propria email istituzionale.

NOTA: per inserire il parametro path di cui al punto a) precedente potrebbe essere necessario espandere la sezione “altre opzioni” al fondo della finestra.



UPLOAD DEI FILE TIFF su Google Cloud Storage

Creare una cartella di lavoro facendo click in un punto bianco con il pulsante destro del mouse e scegliendo “Nuova cartella” e specificando il nome desiderato. Entrare nella cartella facendoci **doppio click sopra**. Nell'esempio che segue la cartella è stata denominata **BEM-PRODUZIONE**



All'interno di questa cartella - che conterrà tutti i progetti della biblioteca - si potranno creare tutte le cartelle e sottocartelle che si desidera per archiviare i propri progetti in modo strutturato.

E' quindi importante creare una sottocartella per ogni progetto e all'interno di ogni cartella di progetto creare un'ulteriore sottocartella per ogni oggetto digitale che si andrà a creare. Questa ulteriore sottocartella dovrà avere un nome che contiene anche il BID dell'oggetto digitale.

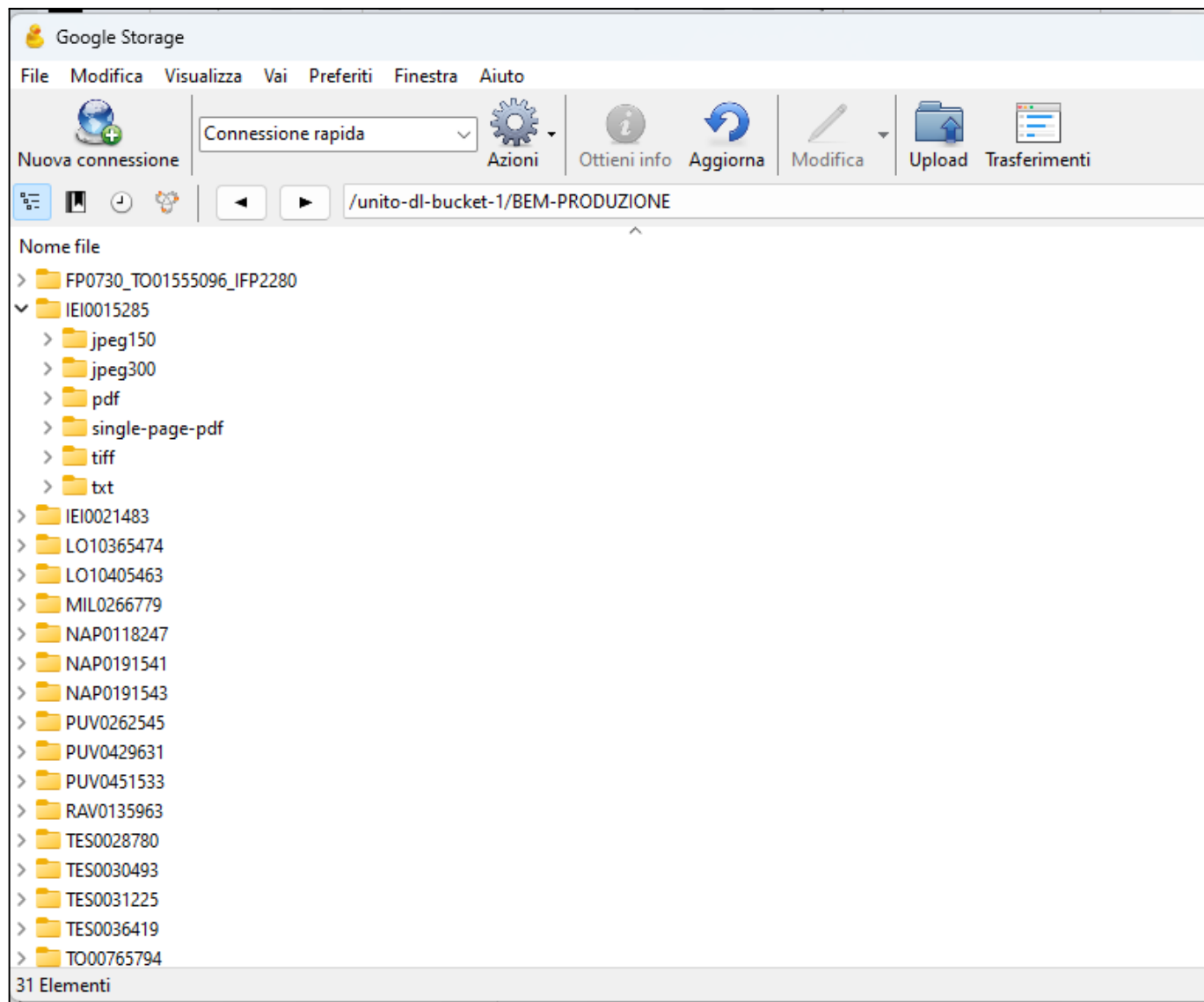
Nel caso della BEM il formato dei nomi sarà

COLLOCAZIONE_BID_INVENTARIO (es. FP0216_LO10353293_IFP3030)

All'interno della cartella appena indicata dovrà essere caricata l'intera cartella “tiff” con i file immagine creati dalla scanner e opportunamente nominati.

ATTENZIONE: fare sempre **doppio click per entrare in una cartella** altrimenti i file che si caricano verranno caricati in luoghi non desiderati.

Nell'immagine che segue si può vedere un esempio della struttura dello spazio BEM su Google Cloud Storage.

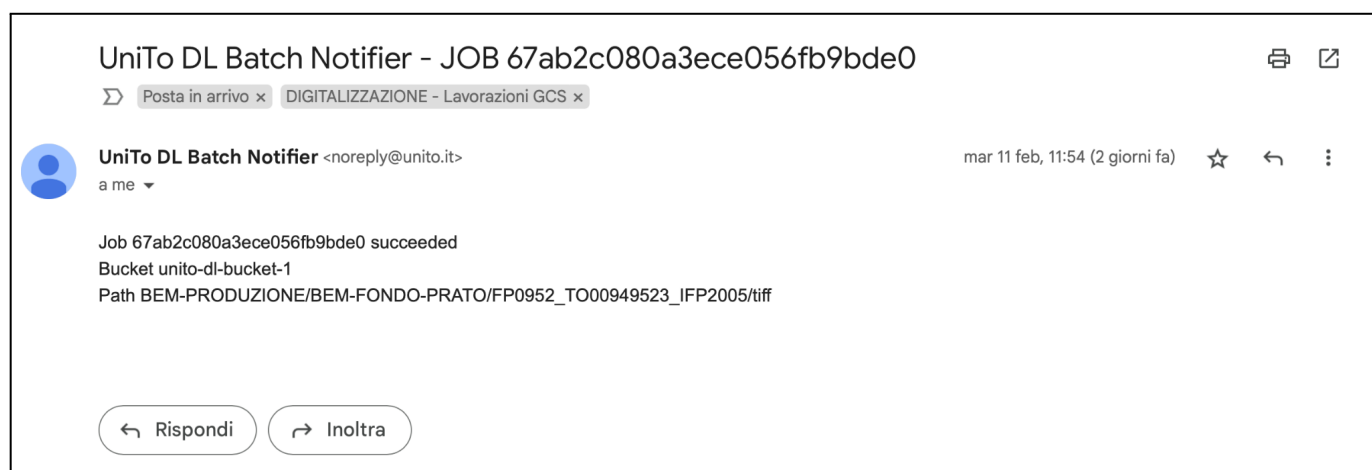


Per fare generare tutto il pacchetto di file in modo automatico da Google Cloud Storage partendo dalle immagini tiff, procedere in questo modo:

1. caricare nella cartella BEM-FONDO-PRATO la cartella con le immagini tiff provenienti dallo scanner
2. **Attendere il completamento dell'upload** e premendo diverse volte sul pulsante "Aggiorna" a forma di circoletto, attendere che tutti i file siano effettivamente presenti su GCS. Il numero di files sul pc deve corrispondere al numero di file su GCS.
3. **Solo a questo punto**, caricare nella cartella contenente le immagini tiff il file di testo **PROCESS.cmd** creato sul proprio PC e contenente l'unica parola "PROCESS" al suo interno

(vedere oltre per creare questo file). Il caricamento di questo file attiva la creazione delle cartelle “jpeg150”, “jpeg300”, “txt”, “pdf” e “single-page-pdf” (*).

4. Attendere che tutte le cartelle siano state create e che in ciascuna vengano creati tutti i relativi files. Il controllo può essere effettuato premendo di tanto in tanto il pulsante “Aggiorna” a forma di circoletto. Tutte le cartelle create avranno terminato l’elaborazione quando dentro ciascuna ci sarà il numero di oggetti originale (=tante tiff ho caricato, tanti jpeg150, tanti jpeg300, tanti txt e tanti pdf singola pagina dovranno esserci nelle rispettive cartelle).
5. Quando GCS termina la creazione di tutto il necessario invierà all’utente che si è loggato la prima volta che la connessione viene impostata una email che lo avvisa di avere terminato. Qui sotto un esempio di notifica:



Terminato tutto questo è possibile scaricare sul proprio PC le cartelle jpeg150, jpeg300, pdf e txt e successivamente caricare il pacchetto, completo di tiff, dal proprio pc sull’area di edit del MAGTOOL con FileZilla e procedere con metadatazione e pubblicazione (vedere la sezione dedicata per le relative istruzioni).

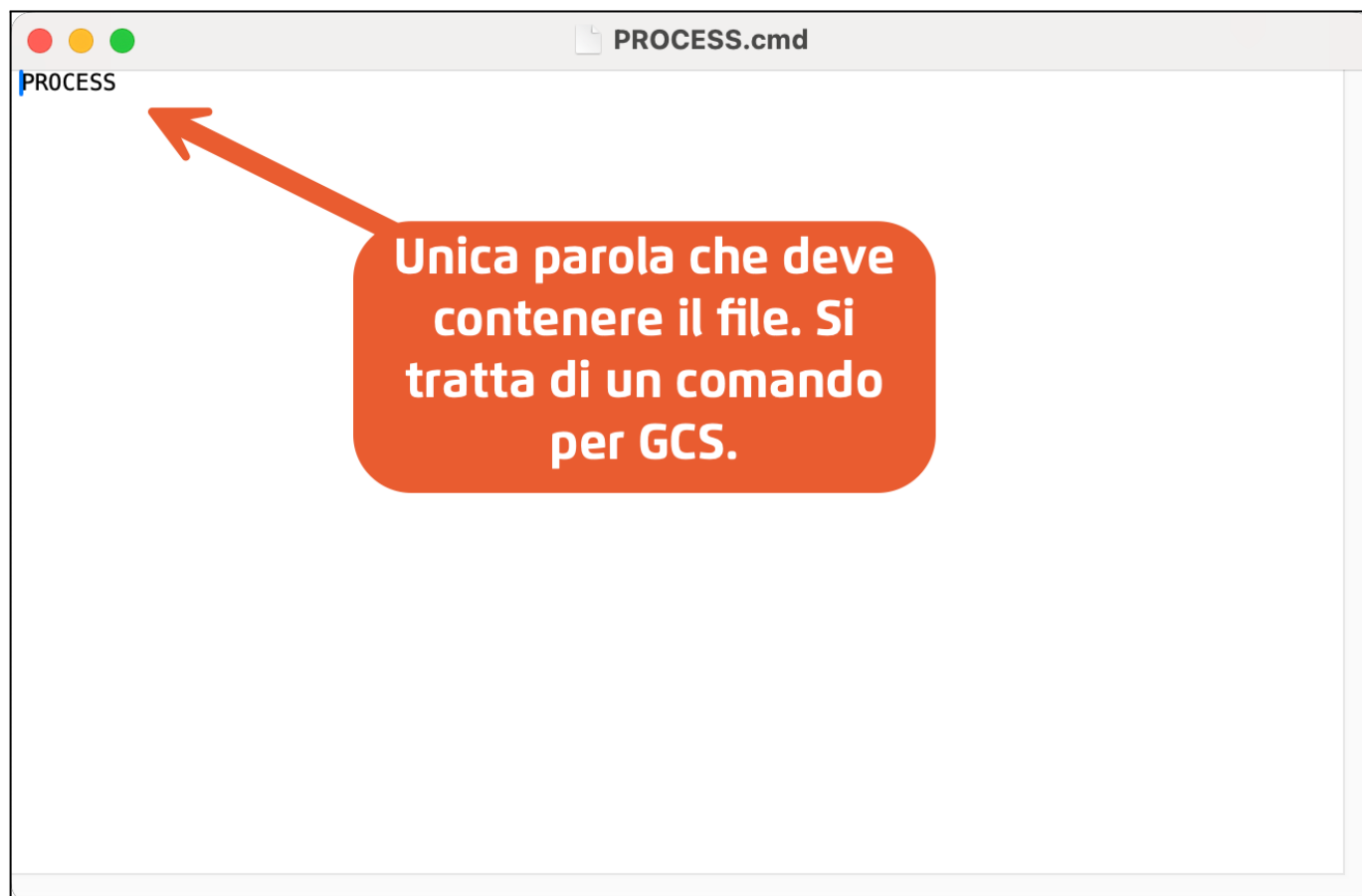
NOTA: non è necessario scaricare sul proprio PC e successivamente sulla edit-area del Magtool la cartella “single-page-pdf” che contiene il pdf delle singole pagine del volume trattato. Questa infatti è solo una cartella di appoggio di GCS per la creazione del file pdf complessivo del volume (che si troverà nella cartella “pdf”).

File COMMAND.cmd

Creare sul proprio PC un file di testo con nome PROCESS (maiuscolo) ed estensione “cmd” (minuscolo) che al suo interno contiene l’unica parola PROCESS (maiuscolo).

Nota: attenzione che il file finale non dovrà avere estensione .txt ma .cmd

Questo file sarà quello da caricare nella cartella “tiff” su Google Cloud Storage.



Creazione dei file (vecchio metodo: senza usare Google Cloud Storage)

Nel caso non si disponga del sistema Google Cloud Storage o questo sia momentaneamente non disponibile è possibile creare il pacchetto di file necessario con il vecchio metodo indicato nel seguito.

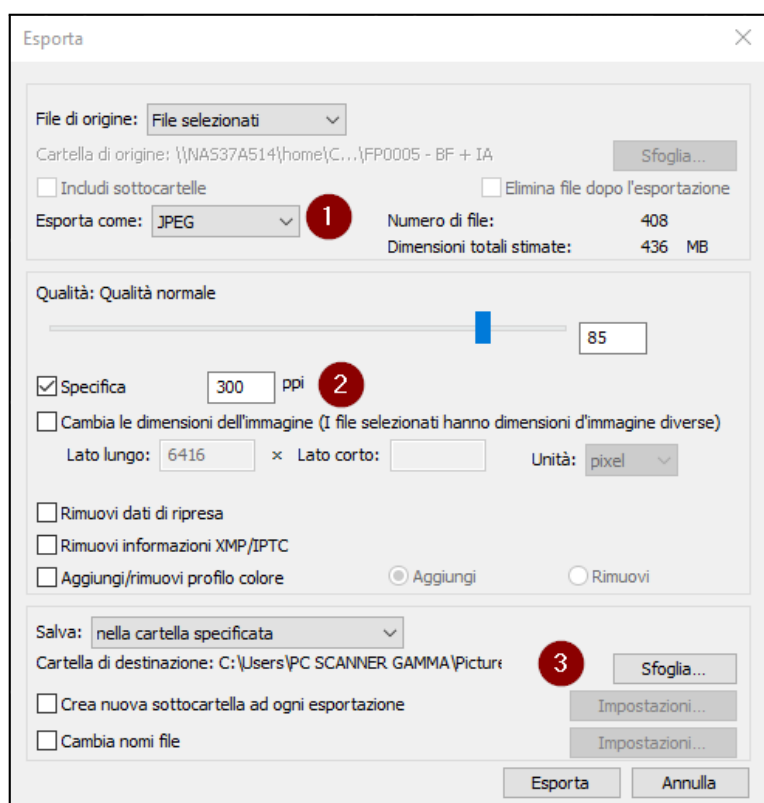
File TIFF

Sono il risultato prodotto dallo scanner durante la scansione (vedere sezione dedicata)

File JPEG (sia a 150 che a 300 dpi):

Questi file vengono creati a partire dai TIFF prodotti dallo scanner usando il software NX Studio della Nikon procedendo come segue (la procedura si riferisce alla creazione di quelli a 150 dpi e andrà ripetuta una seconda volta modificando l'impostazione dei dpi a 300 per gli altri):

1. Aprire il software NX Studio
2. Nella colonna di sinistra navigare nel computer fino a trovare la cartella che contiene i file TIFF sorgente
3. Nella parte centrale, una volta che viene visualizzato il contenuto della cartella selezionare tutti e soli i file tiff delle pagine dell'oggetto digitale numerati da 0001 a nnnn
4. Con i file così selezionati cliccare sul pulsante *Esporta* presente all'estremo in alto a destra nell'interfaccia del programma. Comparirà la finestra seguente dove bisognerà procedere come segue:



- a. In (1) selezionare il formato di esportazione JPEG
- b. In (2) attivare (se non lo è già) la casella di spunta "*Specifica*" e impostare il valore di ppi desiderato (la procedura va ripetuta due volte: una con questa impostazione a 300 e una con 150)
- c. In (3) mediante il pulsante "*Sfoglia*" selezionare la cartella di output sul pc locale dove si vuole andare a salvare tutte le immagini convertite in jpeg.

5. Fare click su “Esporta”
6. Ripetere dal punto 4 una seconda volta per generare le immagini jpeg a 150 ppi.

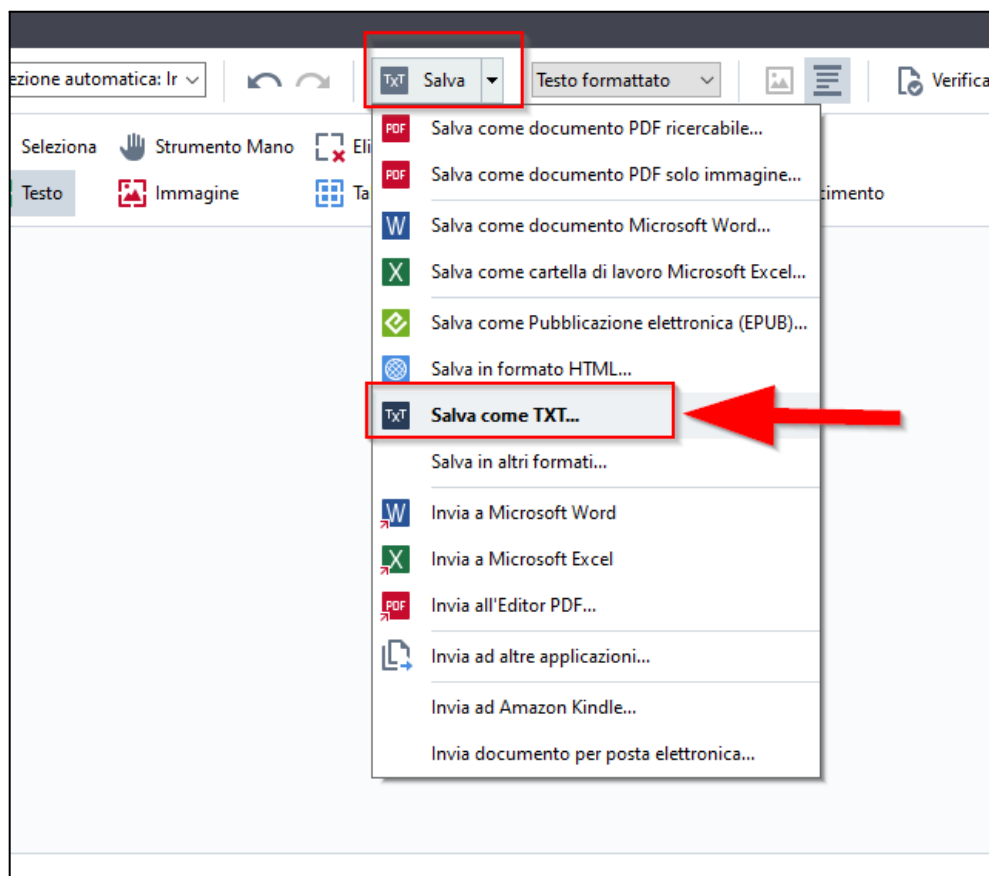
File TXT

Questi file contengono il testo presente nelle immagini tiff (il c.d. OCR o riconoscimento ottico dei caratteri). Per ottenerle si usa il software ABBYY Finereader.

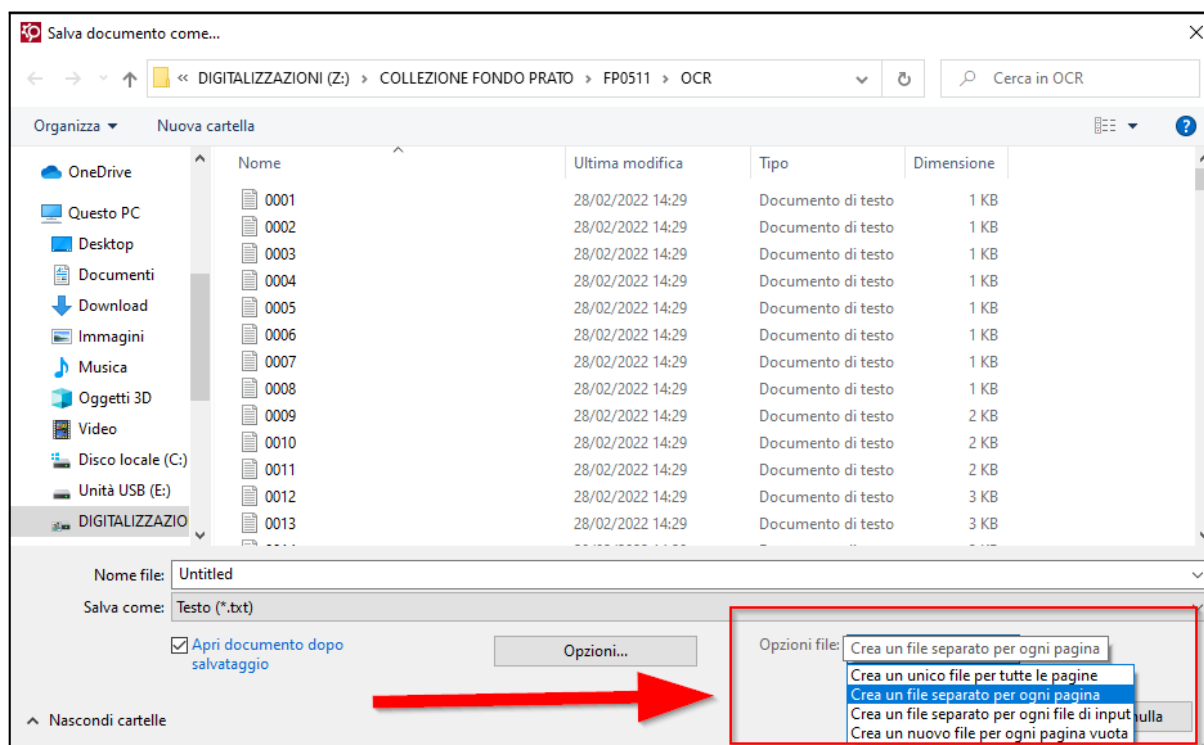
1. Aprire il programma ABBYY Finereader (icona disponibile sul desktop)
2. Cliccare su “Apri nell’Editor OCR” e andare a selezionare sul proprio PC i file sorgente TIFF dell’oggetto che si sta considerando

NOTA: in BEM siamo sempre partiti dai file TIFF ma alcune operatrici e operatori di altri poli sono partiti dai JPEG per le operazioni che seguono.

3. Il programma caricherà tutte le immagini TIFF al suo interno e dopo qualche secondo avrà terminato
4. A questo punto cliccare sulla freccina accanto al pulsante “Salva” presente sulla barra dei menu e scegliere l’opzione “Salva come TXT...”



- Specificare nella finestra che segue la cartella dove andare a salvare tutti i file TXT che verranno prodotti e prima di confermare aprire il menu a tendina “Opzioni file” e scegliere l’opzione “Crea un file separato per ogni pagina”

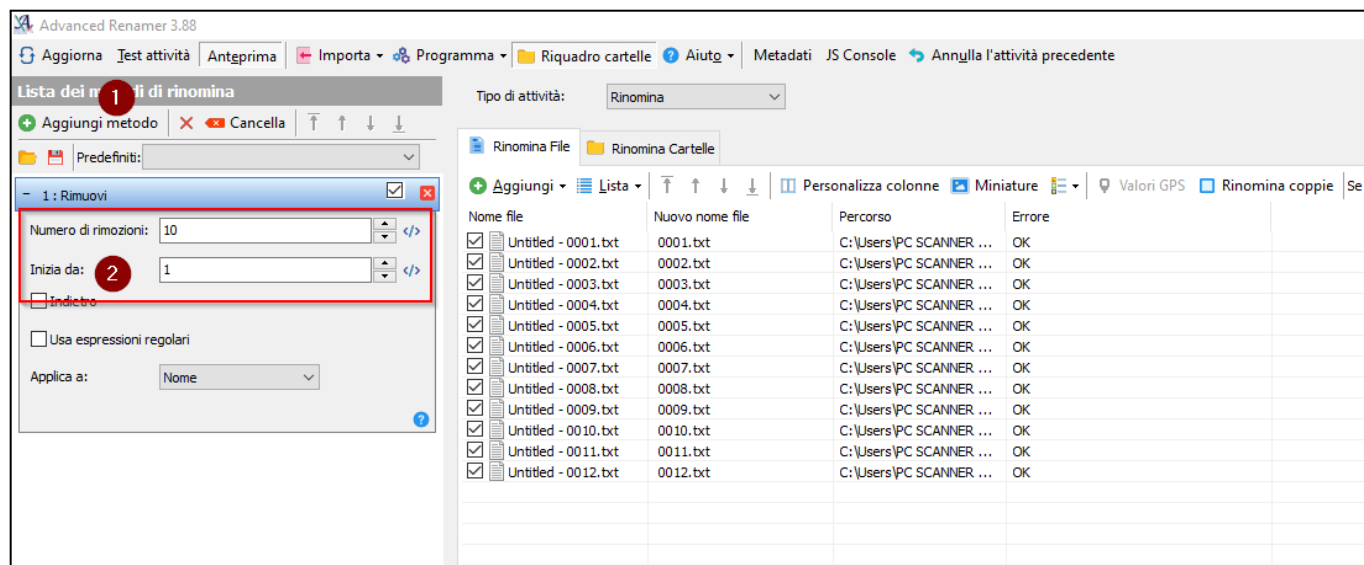


- Confermare con “Salva”. Nella cartella selezionata verranno creati tanti file txt quante sono le immagini tiff con il testo riconosciuto (nota: questi sistemi hanno una certa percentuale di riconoscimento quindi il testo prodotto potrebbe contenere degli errori che potranno essere ignorati).

I file salvati avranno un nome del tipo “Untitled - nnnn.txt”. Si tratta ora di eliminare in blocco dal nome la parte in eccesso in modo che il nome sia nnnn.txt. Per fare questo aprire il programma Advanced Renamer (icona sul desktop o nella barra inferiore di Windows e seguire i passi seguenti:

- Una volta aperto il programma eliminare tutti i metodi di nomenclatura file presenti nella colonna di sinistra (se presenti) cliccando per ciascuno sulla piccola crocetta rossa in corrispondenza
- Cliccare nella parte centrale sul pulsante “Aggiungi → File” e andare a selezionare tutti e soli i file txt creati prima navigando nel PC locale. Questi compariranno nella finestra principale del programma. A questo punto andremo ad applicare a tutti i file un metodo per rinominarli. In questo caso il metodo si chiama “Rimuovi” perchè andremo a rimuovere i 10 caratteri in eccesso presenti a sinistra nel nome del file (es. ~~Untitled~~ - nnnn.txt)

3. Facendo riferimento all'immagine seguente, cliccare dunque su “*Aggiungi metodo*” nella colonna di sinistra (1) e selezionare il metodo Rimuovi. Il metodo comparirà nella colonna già parametrizzato (2). Se non lo fosse impostarlo come nella immagine seguente:



4. Cliccare sul pulsante “Avvio attività” presente in alto a destra nel programma e successivamente su “Rinomina”.
5. Tutti i file saranno rinominati secondo il criterio impostato.

Nome	Ultima modifica	Tipo	Dimensione
 0001	17/06/2022 14:21	Documento di testo	1 KB
 0002	17/06/2022 14:21	Documento di testo	1 KB
 0003	17/06/2022 14:21	Documento di testo	1 KB
 0004	17/06/2022 14:21	Documento di testo	1 KB
 0005	17/06/2022 14:21	Documento di testo	1 KB
 0006	17/06/2022 14:21	Documento di testo	1 KB
 0007	17/06/2022 14:21	Documento di testo	1 KB
 0008	17/06/2022 14:21	Documento di testo	1 KB
 0009	17/06/2022 14:21	Documento di testo	2 KB
 0010	17/06/2022 14:21	Documento di testo	2 KB
 0011	17/06/2022 14:21	Documento di testo	2 KB
 0012	17/06/2022 14:21	Documento di testo	3 KB

Attenzione: la procedura di nomenclatura dei file è irreversibile!

File PDF/A

La creazione del file PDF/A avviene esattamente come fatto per i file txt menzionati sopra. L'unica differenza è che al punto 4 si dovrà scegliere *“Salva come documento PDF ricercabile...”* e al punto 5, nel menù a tendina prima di salvare, si dovrà scegliere l'opzione file denominata *“Crea un unico file per tutte le pagine”*

La creazione dei file necessari e del pacchetto di ingestion è ora terminata. Si ricorda che tutti i file vanno organizzati secondo le indicazioni date nel paragrafo “Cosa bisogna ottenere”.