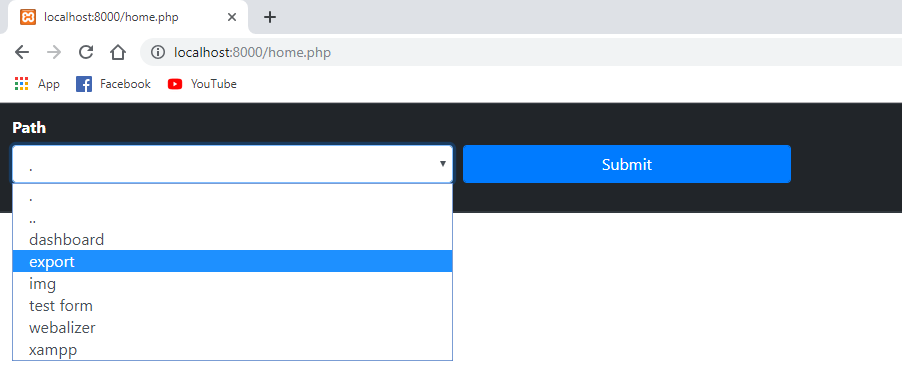
# CARICAMENTO COLLEZIONI NEL SEMANTIC FRAMEWORK: PULIZIA FILE

## 1. Interfaccia di selezione

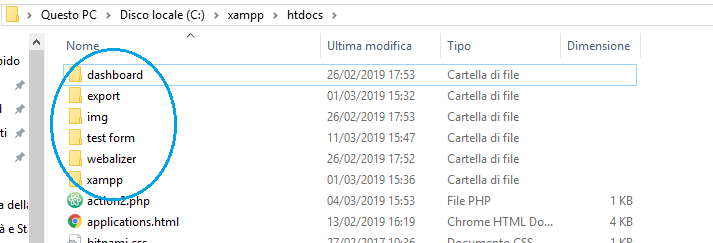
Abbiamo realizzato un servizio web scritto principalmente in php che legge dal percorso specificato i file relativi ai deriverable estratti dalla piattaforma “interregMed”.

Il servizio è stato realizzato allo scopo di rendere più significativi i documenti all’interno della piattaforma.

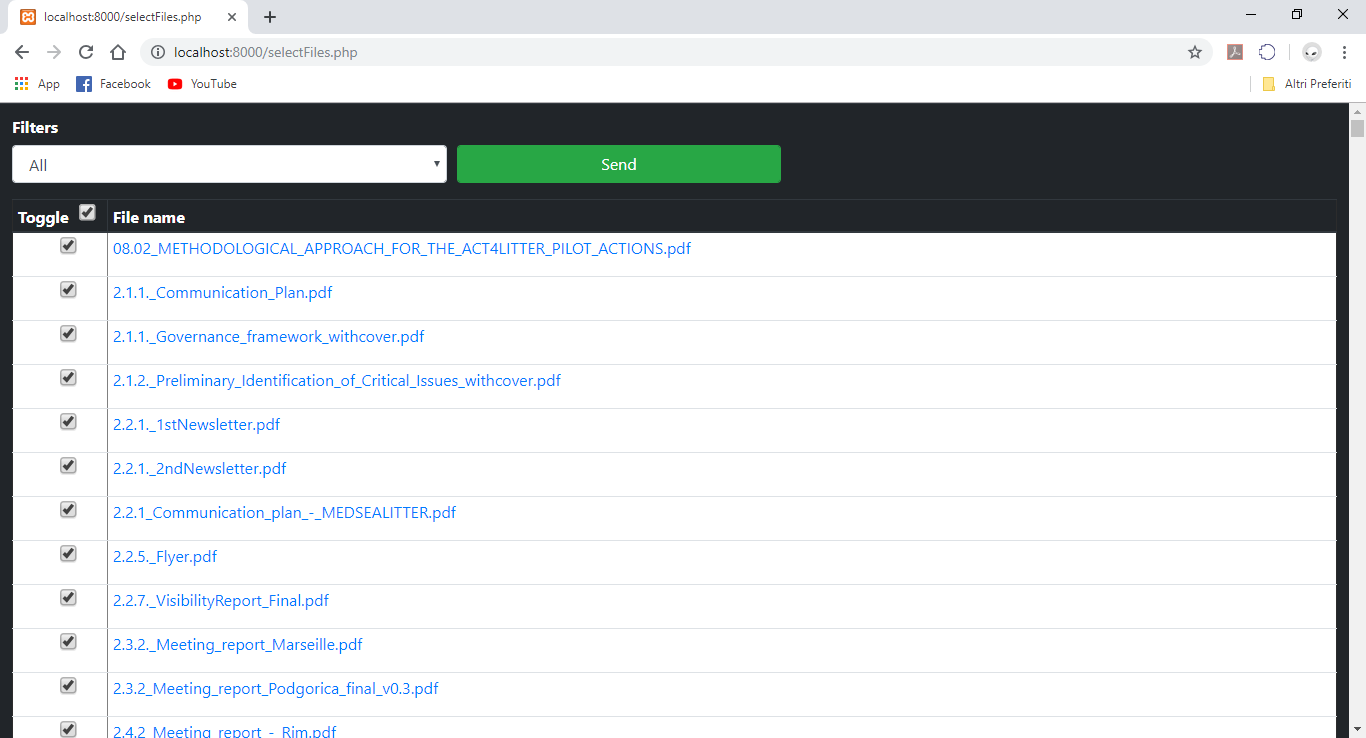
L’interfaccia principale consiste di una schermata in cui poter scegliere il percorso in cui si trovano i documenti (momentaneamente i documenti vengono estratti dalla piattaforma per mezzo di un crawler).



Le cartelle presenti nel percorso impostato nello script vengono quindi mostrate in un menu a tendina ottenendo un risultato di questo tipo. Esempio percorso:



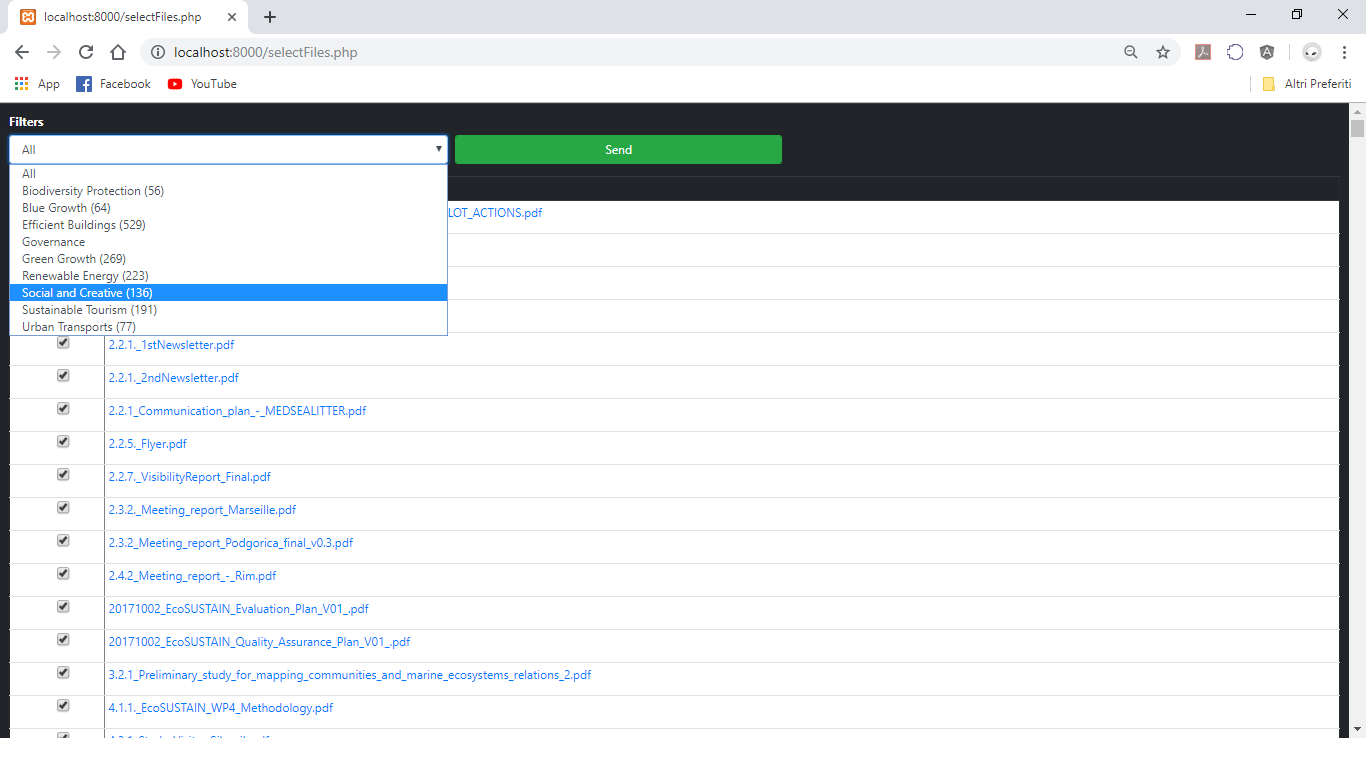
Selezionando ‘export’ e cliccando su ‘Submit’ il sistema esplorerà il percorso specificato visualizzando tutti i documenti al suo interno ottenendo una schermata simile a questa:

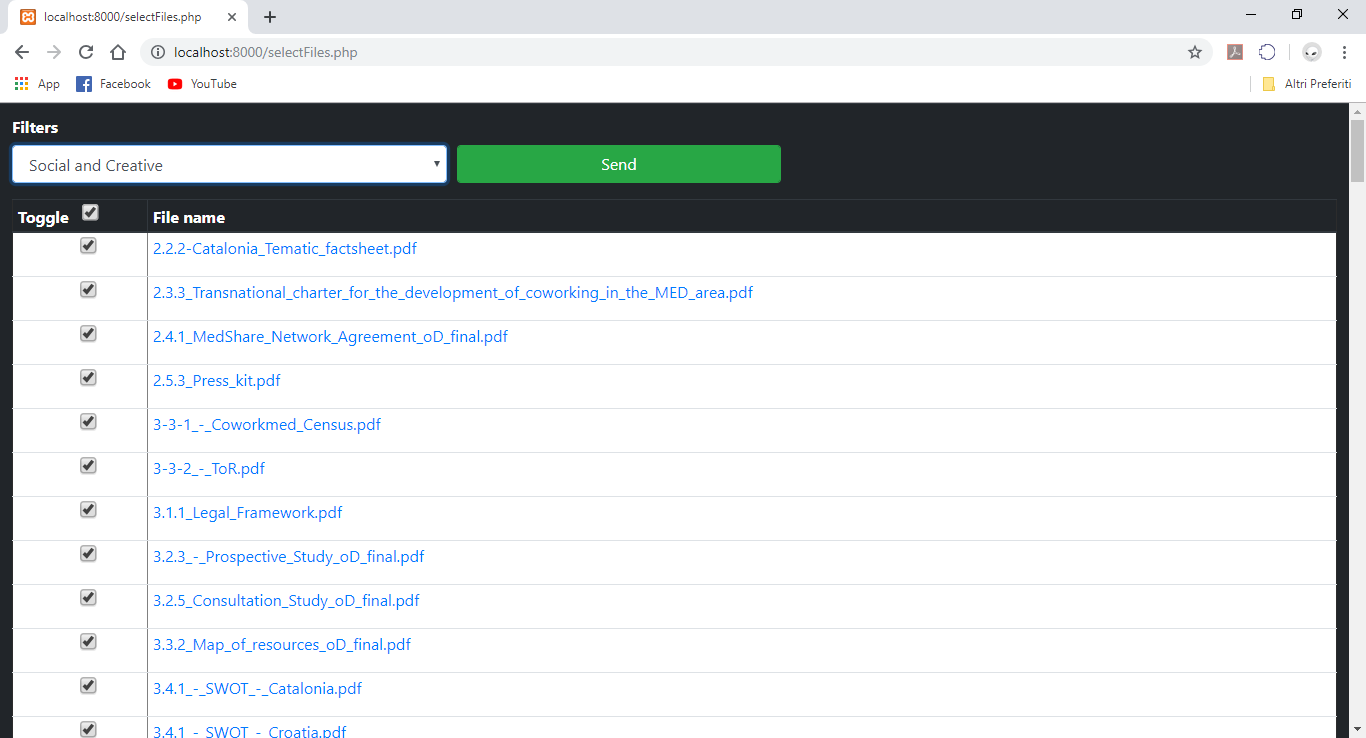


Abbiamo realizzato una struttura di visualizzazione a tabella in cui ad ogni riga corrisponde un documento (deliverable) con annessa casella di selezione per poter scegliere se mantenere il documento (selezionata) oppure scartarlo (deselezionata). I file sono di default selezionati ma è possibile deselezionarli tutti cliccando sulla casella ‘Toggle’ che agirà su tutte le caselle del filtro corrente.

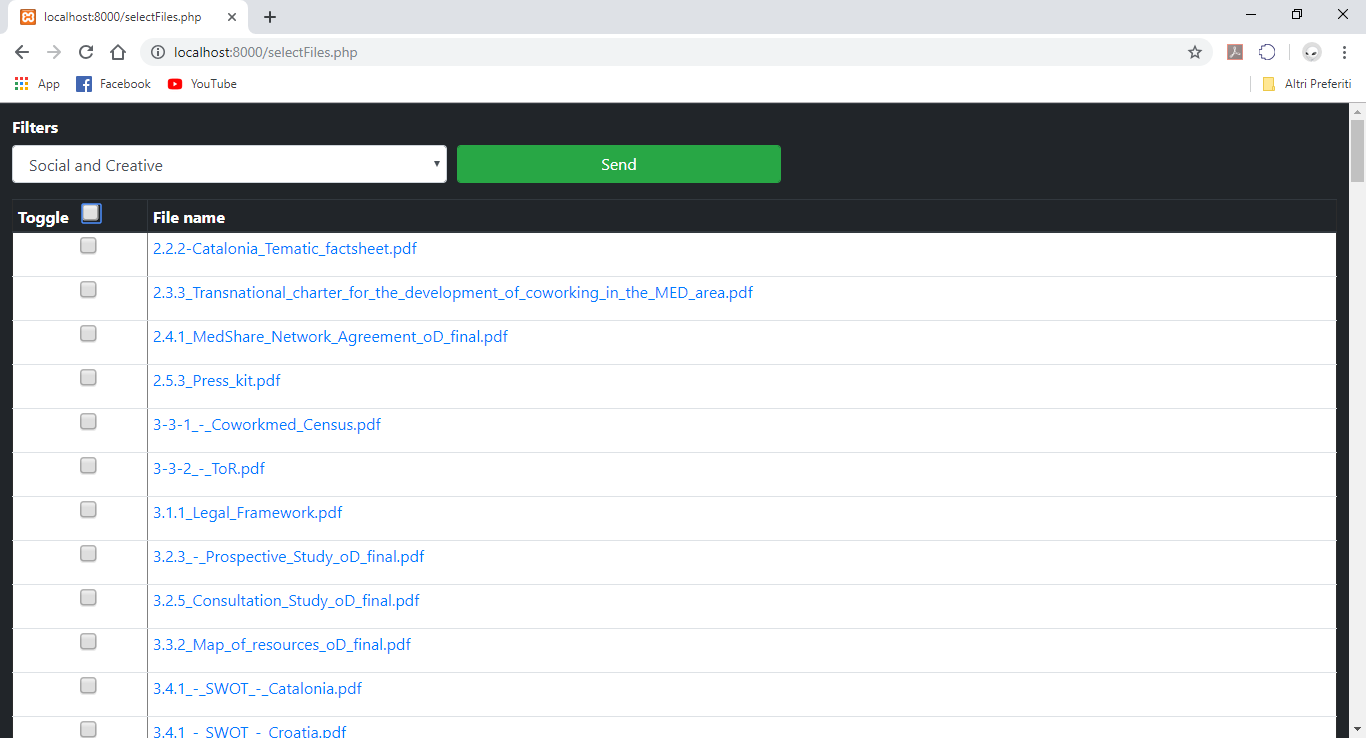
Il nome del file è un collegamento al documento vero e proprio e, cliccandolo, verrà aperta una nuova pagina (non perdendo il lavoro fatto) in cui sarà possibile visionare il documento per scegliere se mantenerlo o scartarlo.

È possibile filtrare i file da visualizzare per collezione di appartenenza utilizzando l’apposita scelta ‘Filters’ e premendo poi sulla collezione desiderata, in questo modo visualizzeremo solo i file presenti nella collezione scelta. Accanto ad ogni collezione viene mostrato il numero di documenti ad essa appartenenti.

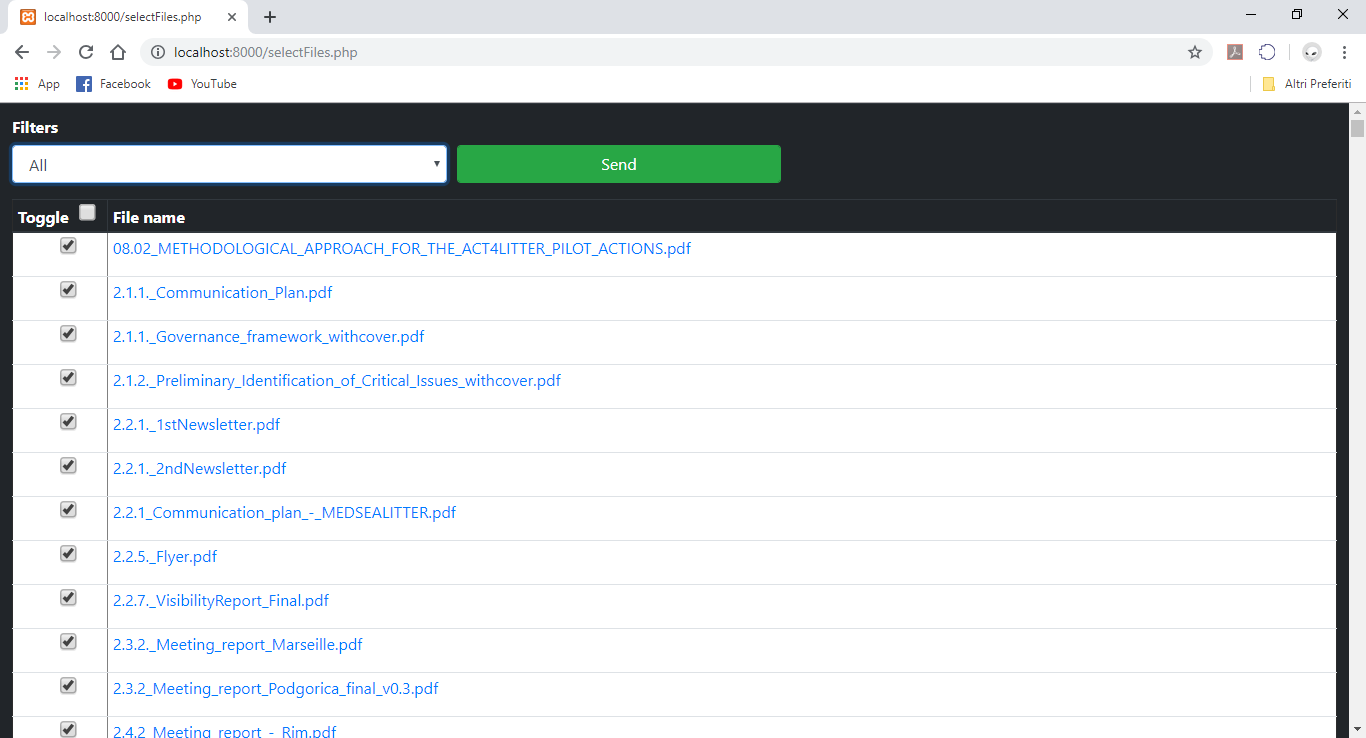
(Esempio: Social and Creative, 136 Documenti)



Ora selezionando la casella ‘Toggle’ si deselezioneranno solo i documenti relativi a ‘Social and Creative’:

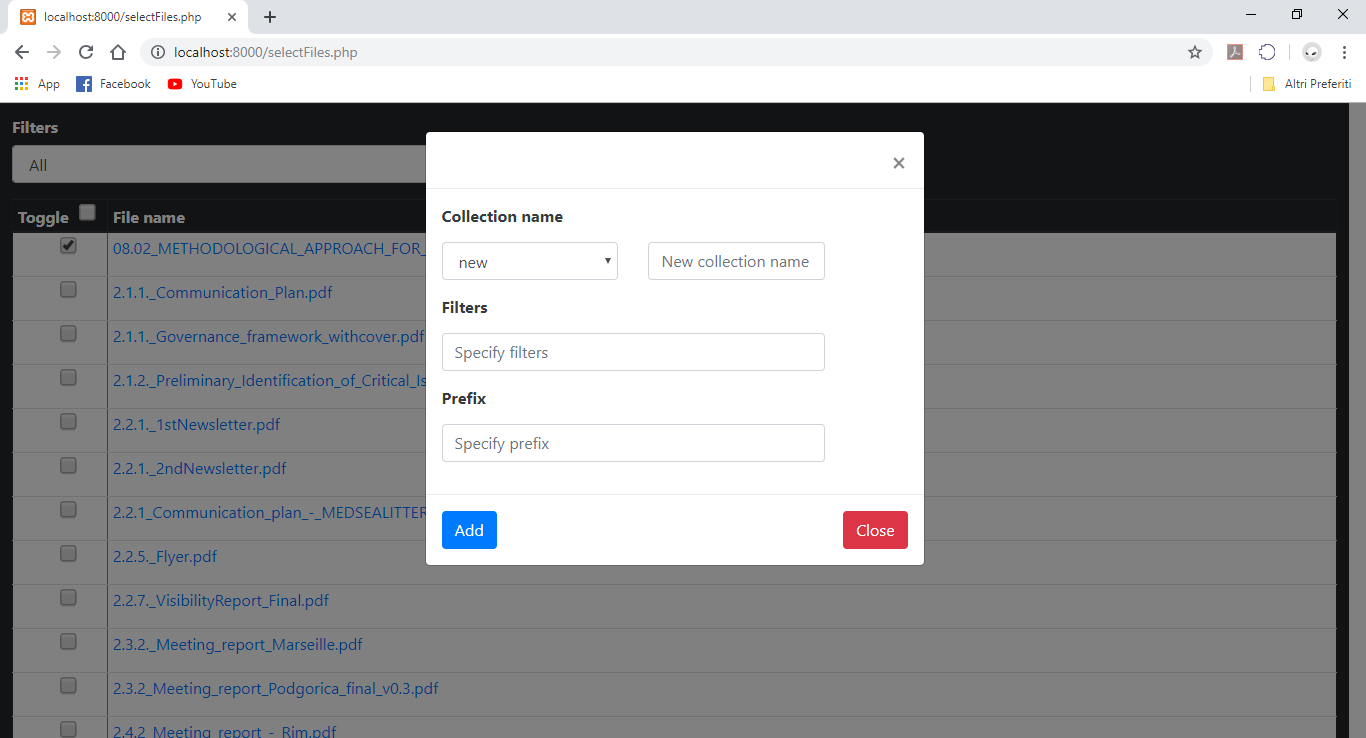


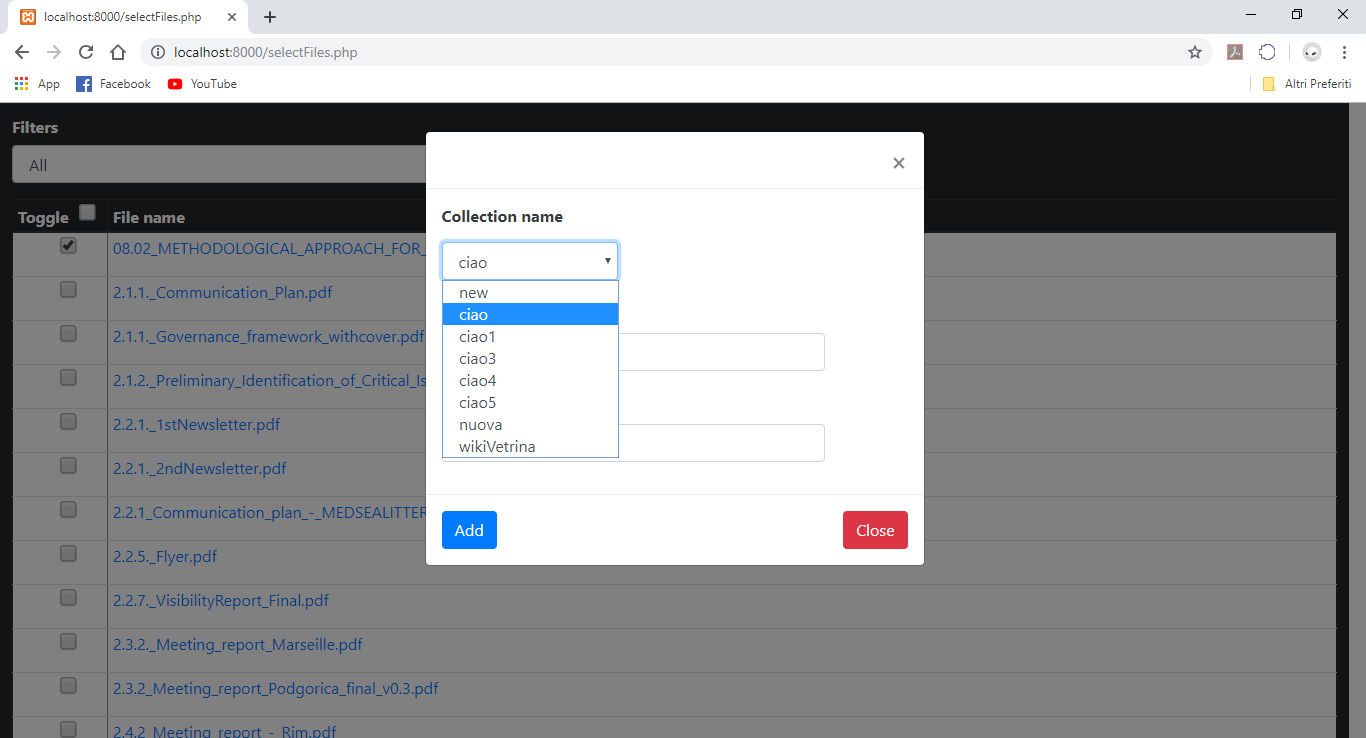
E, premendo sul filtro ‘All’, il filtro verrà resettato (visualizzando tutti i documenti) mantenendo i file deselezionati.



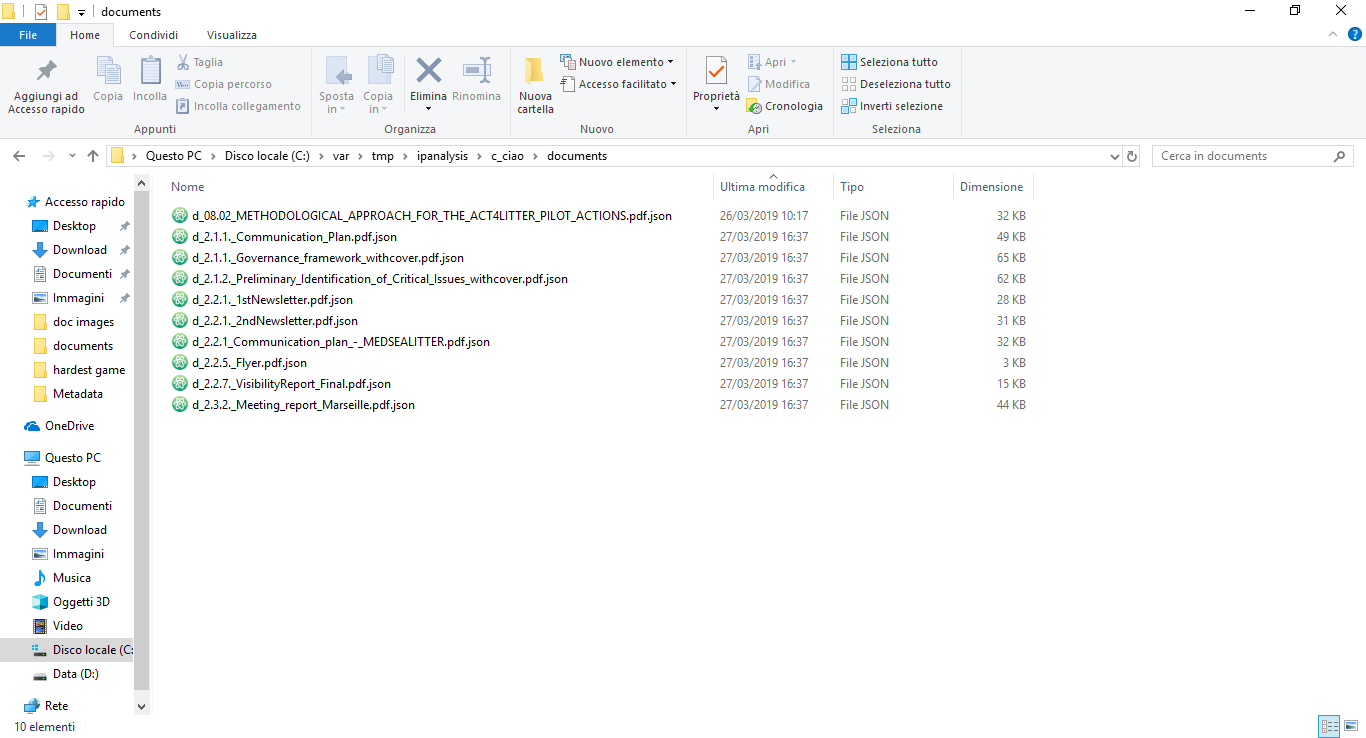
Una volta finito di selezionare/deselezionare i file desiderati, cliccando su ‘Send’ verrà mostrato un popup dove poter inserire le informazioni relative a nome della collezione in cui verranno caricati i documenti (è possibile scegliere tra le collezioni esistenti oppure crearne una nuova, in tal caso bisogna specificare il nome), nome dei filtri (opzionale) e prefisso (opzionale).

## (Esempio primo file caricato)





I file saranno quindi caricati nel framework innovapuglia nella collezione desiderata.



## 2. Elaborazione delle informazioni

Dopo aver selezionato i documenti da caricare e la collezione in cui verranno caricati, verrà effettuata una serie di operazioni sui file al fine di rendere possibile l’estrazione di informazioni aggiuntive.

Per ogni file selezionato, viene innanzitutto applicata la pipeline di NLP (Natural Language Processing) la quale estrae i token, il tag posizione, i lemmi, e parti di frasi.

Una volta eseguito questo per tutti i documenti, verranno eseguite diverse operazioni:

* Estrazione dei bigrammi utilizzando il metodo “Word to Phrases”;
* Analisi della collezione creata con conseguente indicizzazione, estrazione dei concetti e analisi semantica;
* Creazione di un motore di ricerca classico, basato quindi sulla corrispondenza delle parole
* Creazione di un motore di ricerca semantico, basato sulla corrispondenza dei concetti, che utilizza il metodo del Random Indexing