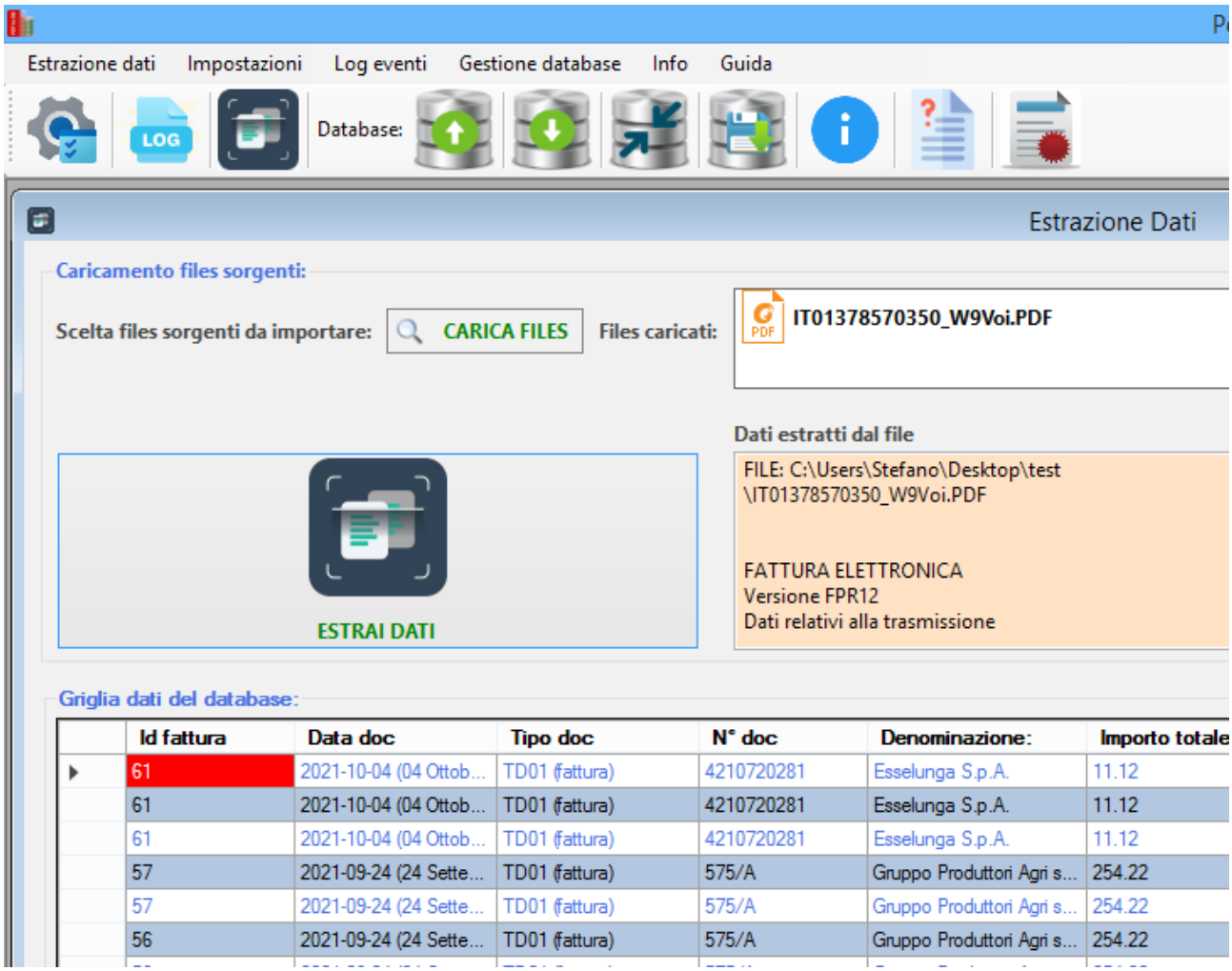
PdfToDb, Estrazione di dati da PDF e salvataggio su database

Sviluppato da Stefano Ravagni V 1.0.0.0 – Ottobre 2021



Indice generale

PdfToDb, Estrazione di dati da PDF e salvataggio su database

Sviluppato da Stefano Ravagni V 1.0.0.0 – Ottobre 2021

PdfToDb, Estrazione di dati da PDF e salvataggio su database

Sviluppato da Stefano Ravagni V 1.0.0.0 – Ottobre 2021

Cos'è PdfToDb

PdfToDb è un software per l'estrazione di dati da files PDF ed il relativo salvataggio su un database (attualmente MsAccess).

I dati estratti dai files sorgente in formato PDF saranno scegli dall'utente attraverso la configurazione di 5 campi obbligatori e 5 campi opzionali derivati (vedi sezioni successive).

Una volta caricati i files PDF sorgente (1 o n files) i dati saranno estratti automaticamente attraverso la semplice pressione di un bottone; da lì saranno automaticamente salvati su un database creando così una base dati solida da utilizzare in futuro per ricerche e quanto altro si desideri.

Come funziona PdfToDb

Il funzionamento è piuttosto elementare…
Una volta configurato a dovere sarà sufficiente compiere due semplici azioni:

1. Cliccare sul tasto sfoglia e caricare da 1 a n files PDF sorgenti da cui estrarre i dati
2. Cliccare sul bottone estrai dati per dare il via al processo di importazione su database dei campi specificati nelle impostazioni

Non ci sono altri passaggi da compiere per usare PdfToDb !!!

Ovviamente va configurato a dovere ed ecco qui allora una breve spiegazione di come confugurarlo al meglio.

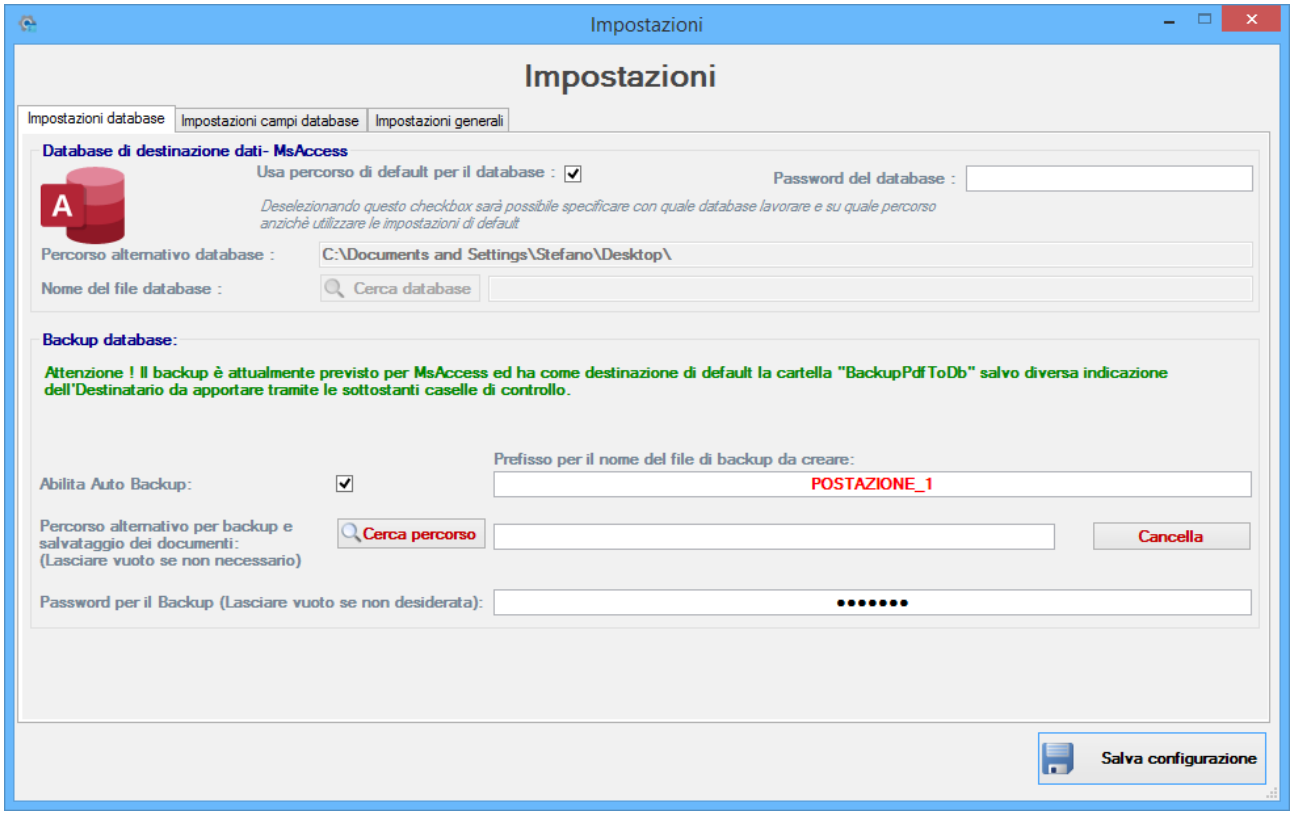
Configurazione di PdfToDb

Le impostazioni di PdfToDb sono raggiungibili dalla toolbar o dal menu relativi alle impostazioni.
La form delle impostazione è suddivisa in 3 schede:

Impostazioni database:

Questa scheda racchiude le impostazioni per l'utilizzo del database Ms Access incluso nel

software.
La scheda è suddivisa in vari comandi ognuno dei quali ha un effetto diverso sul comportamento finale del software.



Uso del percorso di default per il database:

Indica di usare il database incluso nella cartella del programma. Il percorso equivale a quello dove l'utente lo ha installato, solitamente C:\program Files (x86)\PdfToDb .

Disabilitando la spunta sarà possibile indicare dove prelevare un file con estensione .MDB, tipico dei database in formato Ms Access.

Password del dabase:

Qualora il database fosse protetto da password è possibile indicarla al programma in modo che vi possa accedere senza problemi. La password va impostata manualmente tramite Microsoft Access o software in grado di alterare il database.

Abilita auto-backup:

Indica al programma di eseguire automaticamente una copia di backup ad ogni chiusura del software stesso, senza intervento dell'utente.

Prefisso per il nome del file di backup da creare:

Indica una stringa alfanumerica da anteporre al nome del file di backup che sarà creato

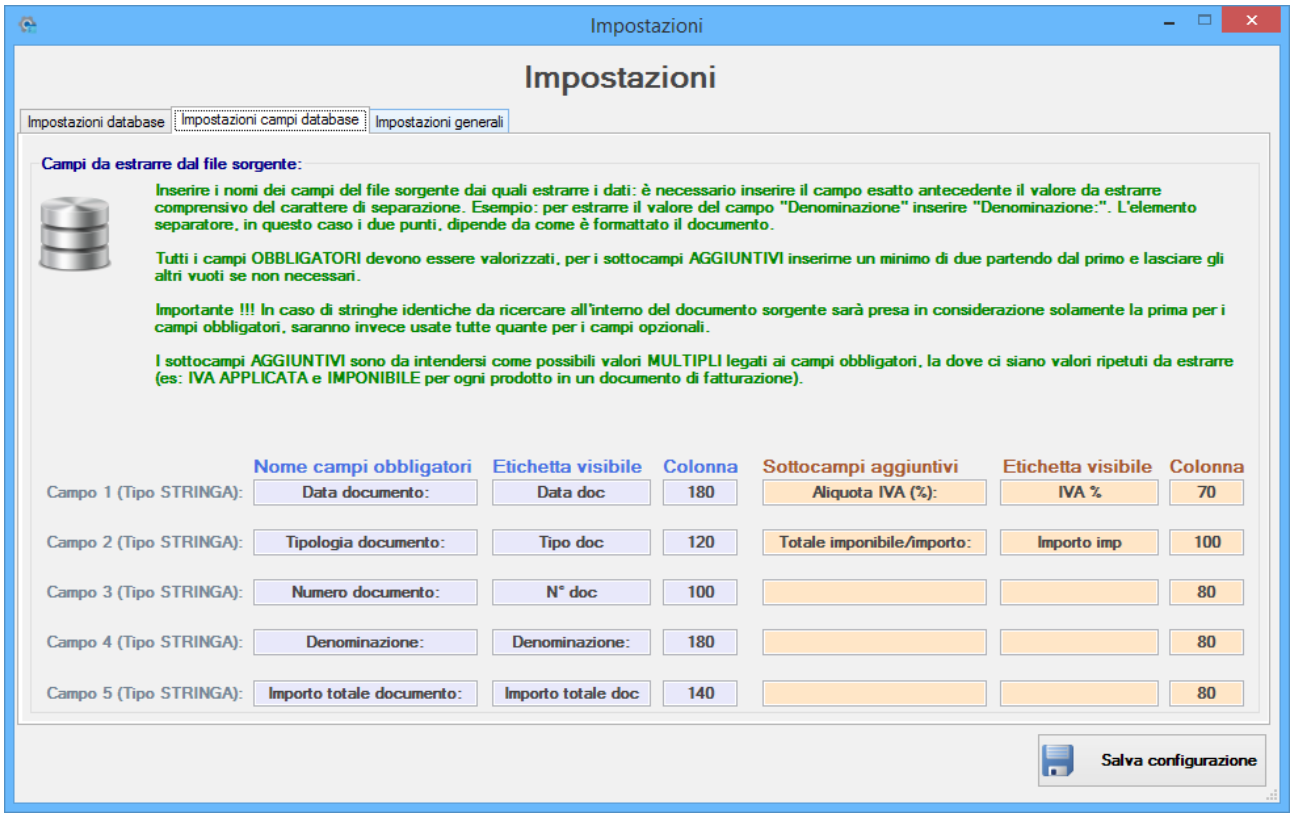
Percorso alternativo per backup e salvataggio dei documenti:

Di default il software salva i backup all'interno della cartella BACKUP della cartella principale dove è stato installato.
Con questo campo è possibile specificare una destinazione diversa per il backup automatico che viene generato. Utile per chi si dimentica di fare un salvataggio manuale frequente che potrebbe scongiurare la perdita accidentale di dati.

Password per il backup:

Indica al programma di proteggere il backup creato con una password specificata dall'utente.

Impostazioni campi database:



PdfToDb permette di cercare all'interno dei files sorgenti in PDF **5 campi fissi** e 5 **sottocampi aggiuntivi**, tutti di tipo stringa alfanumerica.
Per farlo ha bisogno che i dati da estrarre siano disposti ognuno su righe diverse, requisito essenziale.
Ha bisogno inoltre di sapere cosa cercare per ogni riga, ovvero il campo che identifica il dato da estrarre; per questo è necessario specificare alcuni dati per ognuno dei 10 campi da ricercare ed estrarre.

Una volta caricato un file pdf l'utente può scegliere quali campi estrarre dal documento originale, indicandone la stringa esatta nel campo **NOME CAMPI OBBLIGATORI**.
Volendo ricercare ad esempio il valore che si trova dopo la stringa "**Data documento:"** sarà sufficiente scrivere tale stringa nella prima colonna del campo 1; da notare che l**a stringa deve essere esatta e comprensiva del delimitatore**, in questo caso i due punti (":").

Il programma cercherà una riga all'interno del documento e qualora contenesse la stringa "Data documento:" come corrispondenza esatta andrà a leggere i dati che vi sono di seguito per quella riga e li estrarrà come dati da salvare.

Per fare un esempio pratico, ecco cosa succederebbe se venisse trovata la seguente riga….

Data documento: 27/04/2021 (27 Aprile 2021)

Dati di intercettazione = "Data documento:"
Dati estratti = **27/04/2021 (27 Aprile 2021)**

Si fa presente che per i 5 campi obbligatori, **ognuna delle stringhe inserite verrà cercata una sola volta**; in caso di presenza multipla della stessa stringa sarà presa in considerazione soltanto la prima… pertanto si suggerisce di utilizzare **campi univoci** per le estrazioni.

Nella colonna **ETICHETTA VISIBILE** andrà impostato il nome del campo così come lo si vuole vedere nella colonna della griglia dati. Il testo inserito può essere diverso e generalmente si usano stringhe più corte rispetto a quella del campo originario.
Nella colonna **COLONNA** o **SPAZIO** va indicato un numero intero che specifichi la dimensione orizzontale della colonna; questa impostazione è utile per fare in modo che tutte le colonne abbiano la spaziatura corretta in base all'etichetta scelta e al tipo di dati estratto al fine di ottenere un griglia finale ordinata e che dia la possibilità di visualizzare tutto il necessario senza sprecare spazio.

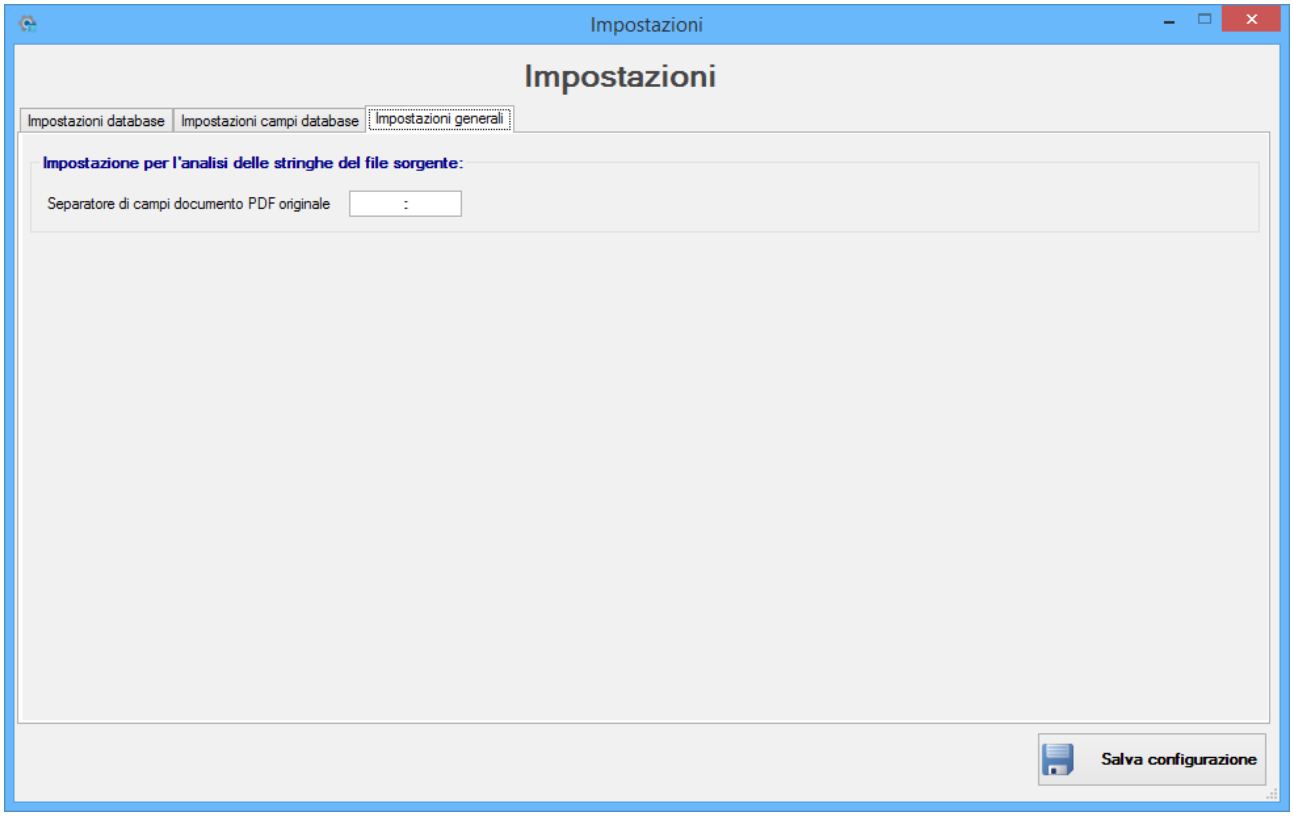
Lo stesso lavoro di configurazione va fatto per tutti i **5 sottocampi aggiuntivi**.

Per i campi aggiuntivi vale lo stesso concetto ma va tenuto presente che questi identificano campi multipli da estrarre quando tutti quanti sono collegati in qualche modo allo stesso documento… per esempio come nel caso delle fatture elettroniche, dove accanto a campi fondamentali come Data tipologie e numero del documento, denominazione etc etc sono presenti campi ripetuti riferiti ai prodotti della fattura stessa… quindi potrò avere 1 solo prodotto o tanti prodotti con aliquota iva e imponibile totale da estrarre… tutti quanti dovranno essere in qualche modo collegati alla fattura nel suo insieme… per questo si parla di sottocampi aggiuntivi multipli.

Devono essere inseriti almeno 2 sottocampi aggiuntivi mentre i restanti 3 sono opzionali.

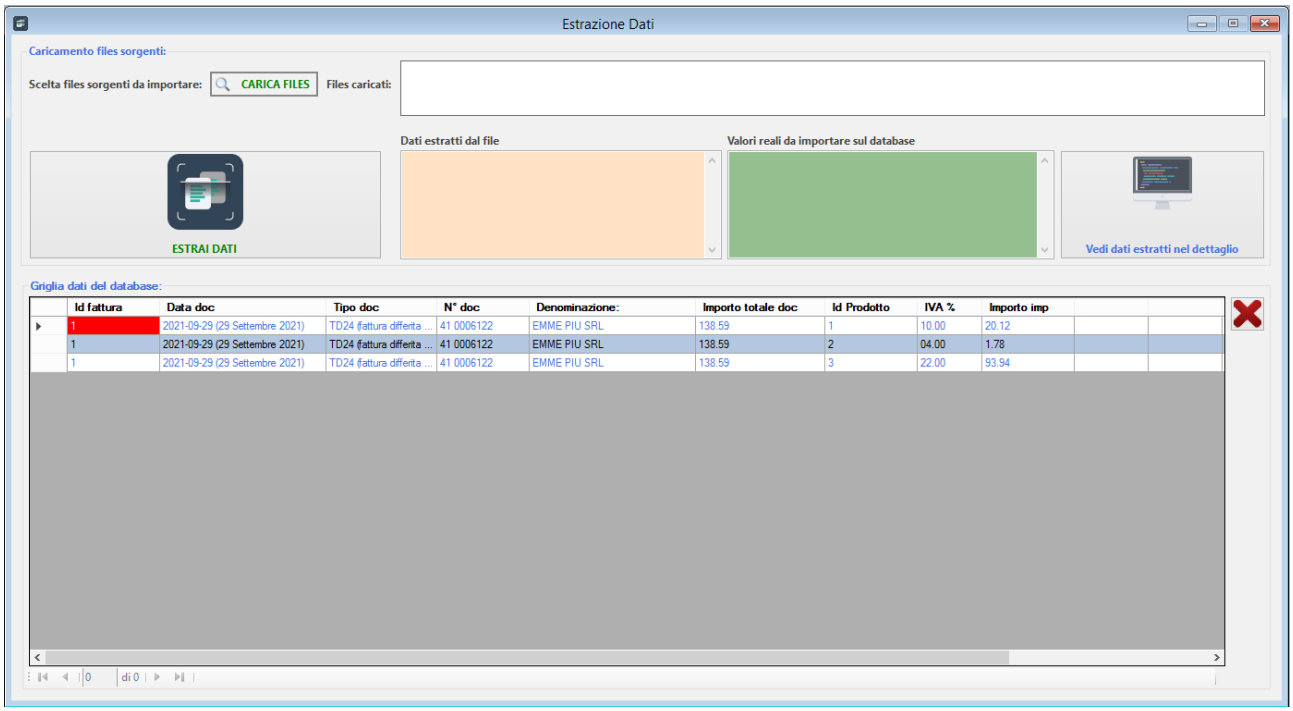
Una impostazioni di questo tipo per i campi obbligatori e quelli opzionali è interessante in vista di futuri cambiamenti nella interpretazione dei documenti… qualora in una fattura elettronica dovesse scomparire il campo DENOMINAZIONE: in favore di COMMITTENTE: tanto per fare un esempio sarà un gioco da ragazzi modificarlo in totale autonomia e continuare ad utilizzare il programma.
La presenza di 5 sottocampi dovrebbe inoltre garantire elasticità per impieghi futuri.

Impostazioni generali:



Tra le impostazioni generali è presente un solo campo per il delimitatore dei campi ma attualmente è **non utilizzato**.

Estrazione dati

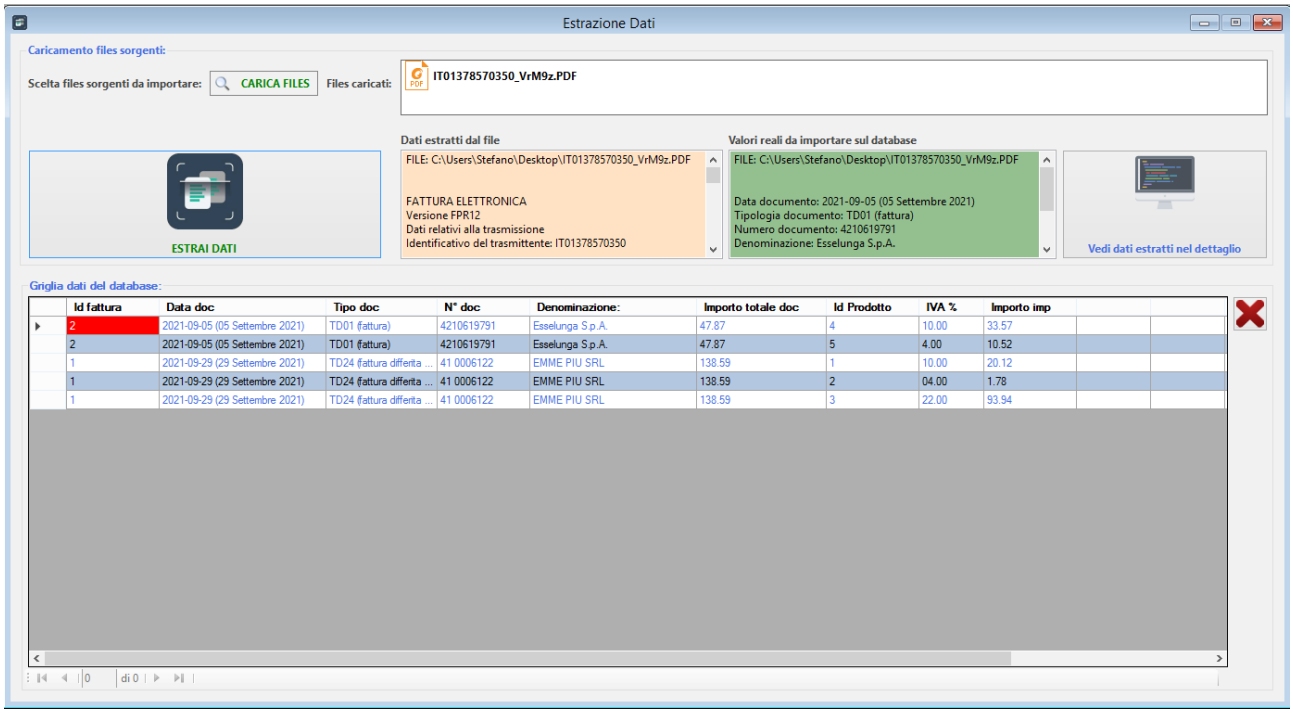


La schermata per l'estrazione dati è la schermata principale del programma PdfToDb ed esegue tutte le funzioni primarie.

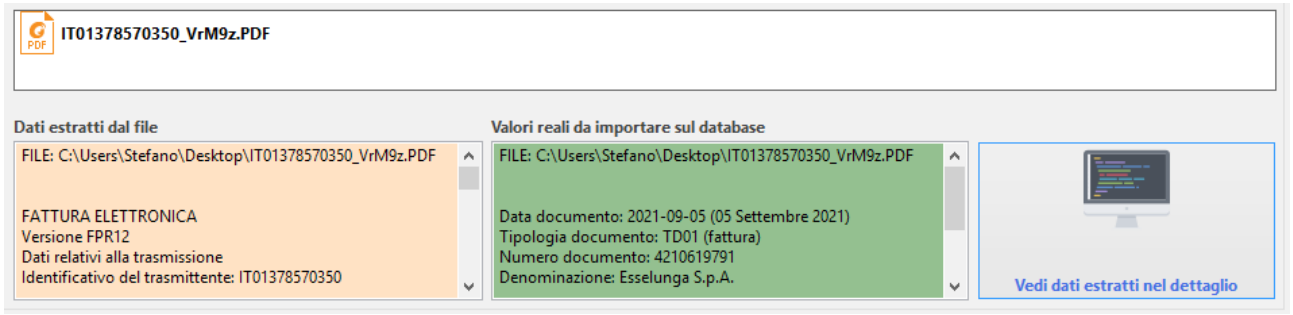
Gli step logici sono i seguenti:

1. Cliccare sul bottone CARICA FILES: questo permetterà il caricamento di 1 o n files sorgenti in formato PDF dai quali estrarre i dati
2. Una volta caricati i dati basterà cliccare sul bottone ESTRAI DATI per dare il via alla estrazione vera e propria.

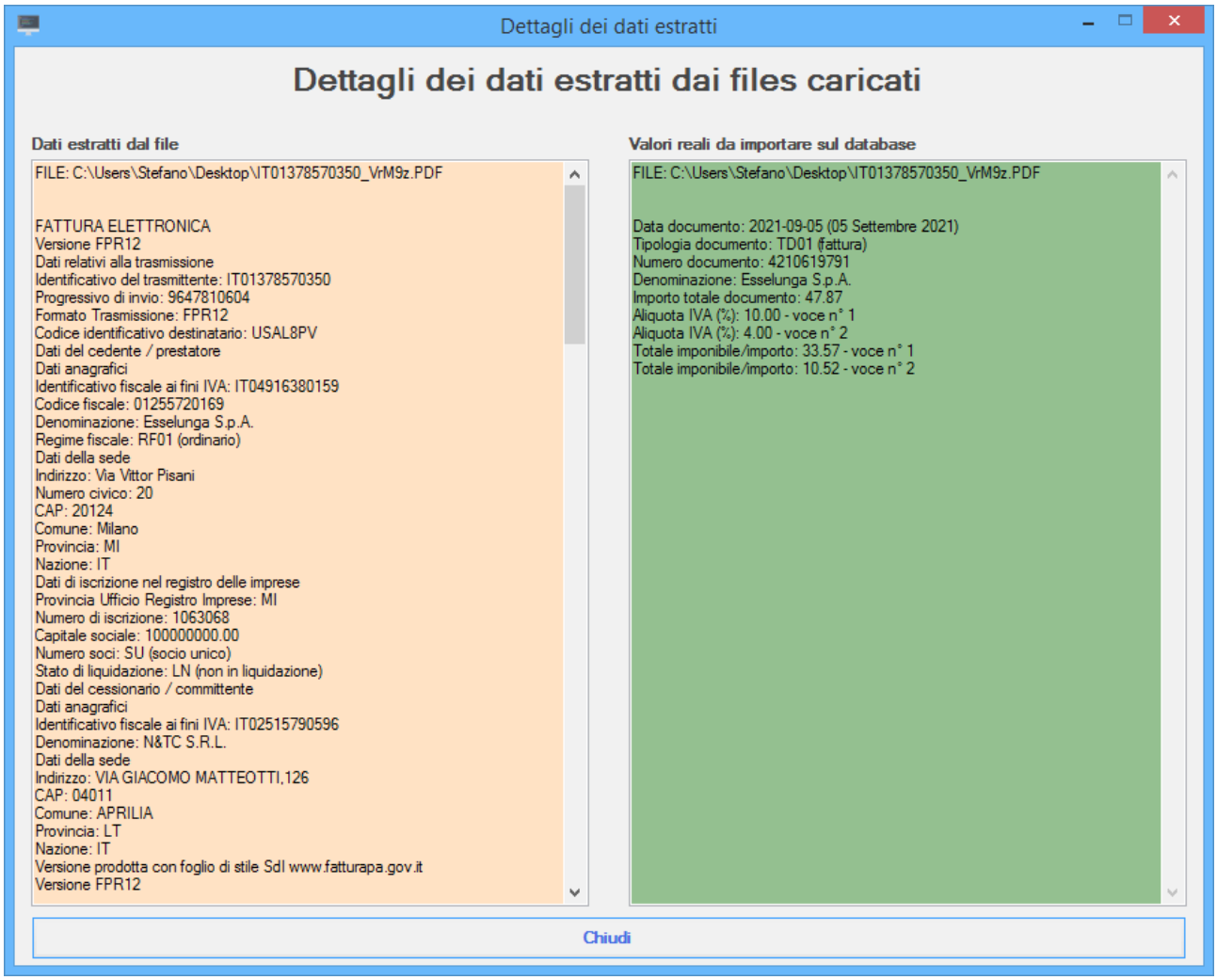
Se l'estrazione è andata bene e non ci sono stati errori ecco quello che apparirà nella form principale.



Sul box DATI ESTRATTI DAL FILE compariranno tutte le righe di ogni files caricato.
Sul box VALORI REALI DA IMPORTARE NEL DATABASE saranno invece riportati i soli dati scegli dall'utente che saranno già stati salvati sul database in base alle impostazioni.



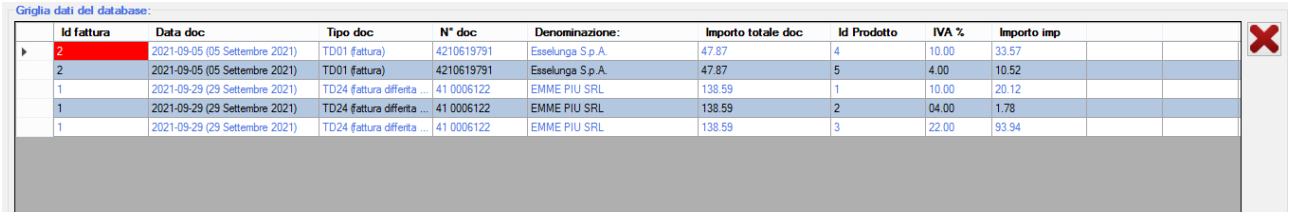
E' possibile vedere questi dati in dettaglio o scorrendo i singoli box o cliccando sul bottone VEDI DATI ESTRATTI NEL DETTAGLIO.



La griglia dati

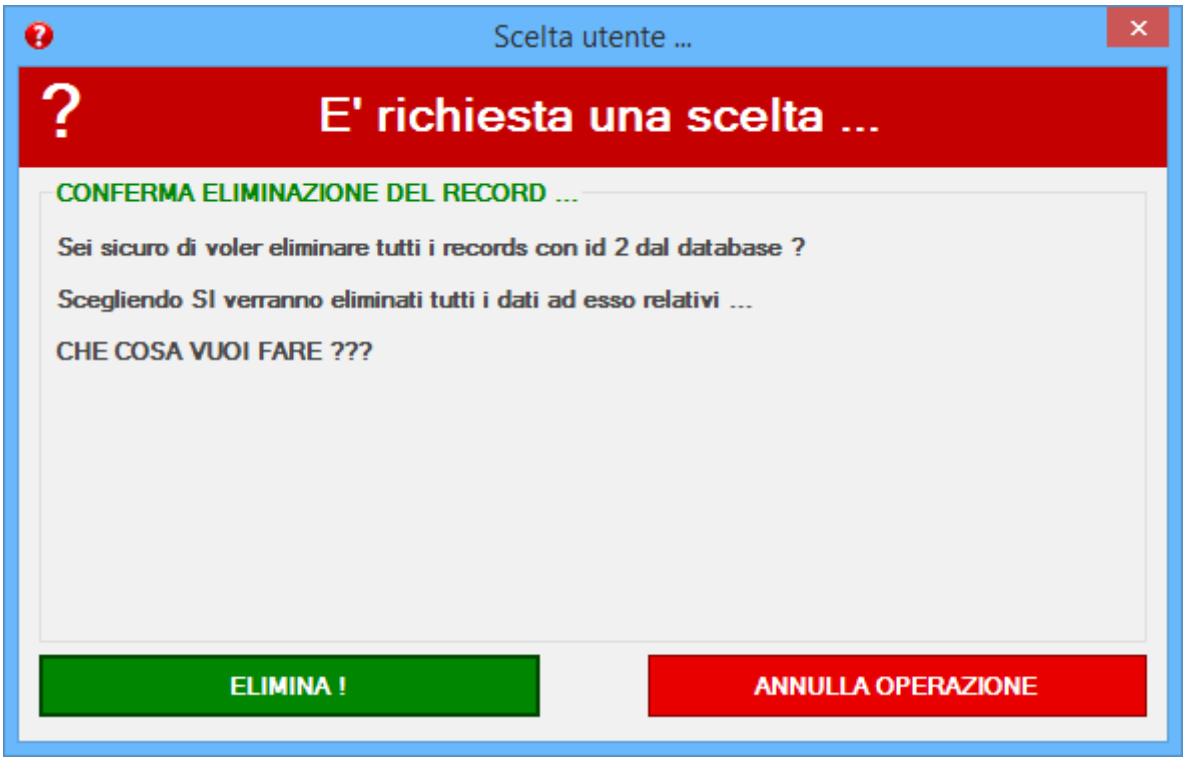
Nella griglia dati saranno visualizzati i dati salvati sul database.

Tutti i dati sono interpretati come dati in formato stringa, in modo da permettere l'estrazione di qualsiasi campo senza limitazioni dovute al tipo di dato.
Questo potrebbe comportare difficoltà nell'utilizzo manuale di query SQL, difficoltà che comunque può essere aggirata tramite apposite funzioni di conversione dei tipi di dato ricercati.



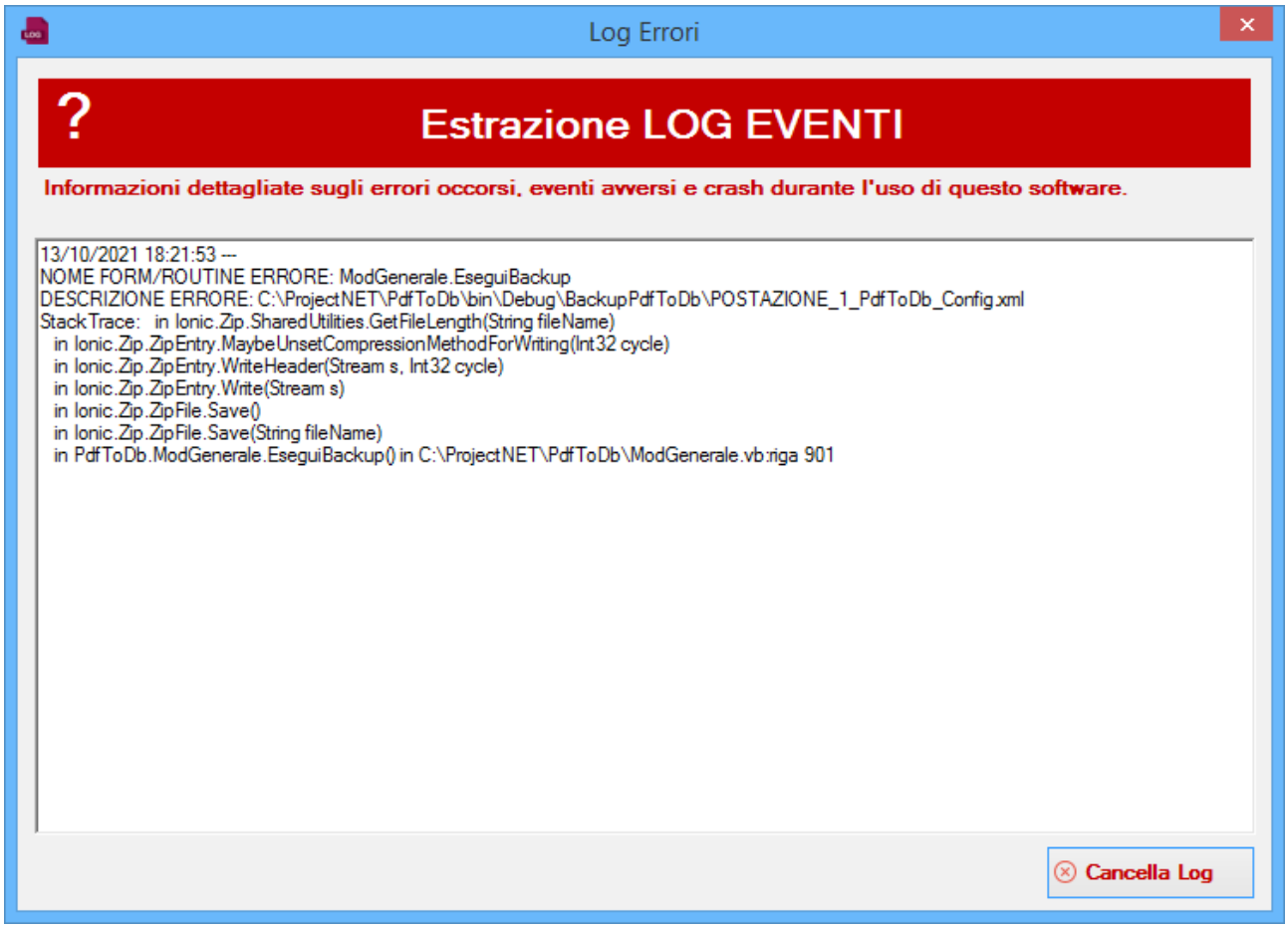
L'ordinamento della griglia è discendente, in alto si trovano le ultime acquisizioni.

Sulla destra è disponibile un bottone per la cancellazione delle righe selezionate; questo può essere molto utile nel caso in cui un documento venga erroneamente acquisito due volte o ci sia stato un errore o ancora non sia più necessario mantenerlo sul database. Prima della cancellazione effettiva l'utente viene avvertito da un popup che chiede conferma dell'azione di cancellazione.



Si fa presente che la cancellazione include tutte le righe che fanno riferimento allo stesso documento originale caricato… nell'esempio già usato delle fatture elettroniche, cancellare una riga che fa riferimento alla fattura numero 10 significa cancellare tutte le righe che fanno anch'esse riferimento alla stessa fattura. Il campo primario è quindi ID FATTURA, ripetuto in fondo alla griglia nell'ultima colonna con riferimento ID FATTURA COLLEGATA.

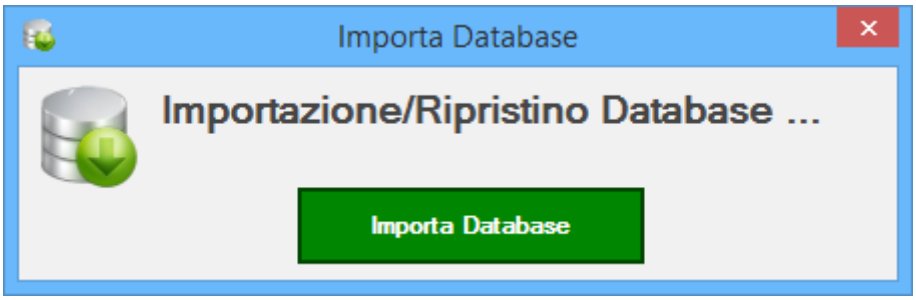
Log eventi



Nella schermata del log eventi è possibile vedere tutti i tracciamenti degli errori occorsi durante l'uso del software. Questi errori possono essere utili a comprendere i motivi di un blocco durante l'utilizzo e dovranno sempre essere comunicati allo sviluppatore per ricevere assistenza.
Si noti che ogni errore tracciato è associato ad un preciso evento, identificato da una data ed un orario preciso.

L'utente può cancellare il file di log con la semplice pressione del bottone CANCELLA LOG.

Importazione database



L'importazione del database è una funzione utile per ripristinare una base dati d un file di backup (salvataggio)… la procedura è guidata e l'utente viene assistito nella ricerca del file database di access da importare o per sovrascrivere la base dati o per ripristinarla da un salvataggio di cui è in possesso.

Esportazione database

L'esportazione del database consente il salvataggio della base dati in un percorso a scelta dell'utente. La procedura è totalmente guidata e l'utente non deve fare altro che scegliere dove creare il backup e come chiamarlo.

Caratteristiche tecniche

PdfToDb necessita della presenza della libreria Microsoft .NET versione 4.8 o superiore. PdfToDb verifica la presenza di questo pacchetto in fase di installazione.

Per chi avesse necessità di installarlo sappia che le librerie di RUNTIME, quelle che ti saranno necessarie, possono essere scaricate al seguente URL:

https://dotnet.microsoft.com/en-us/download/dotnet-framework/net48

Altre librerie necessarie al corretto funzionamento e già incluse nel pacchetto di installazione sono le seguenti:

- -) **iTextSharp .NET PDF library** (Open Source)
  Sito web: http://itextpdf.com/ oppure http://sourceforge.net/projects/itextsharp/

Licenza d'uso (EULA)

La licenza d'uso è inclusa nel pacchetto e mostrata durante la fase di installazione.
L'utente che conclude l'installazione accetta in toto quanto descritto nella licenza d'uso.

Ringraziamenti

Si ringrazia ***Francesco Santopaolo*** per l'idea alla base di questo software