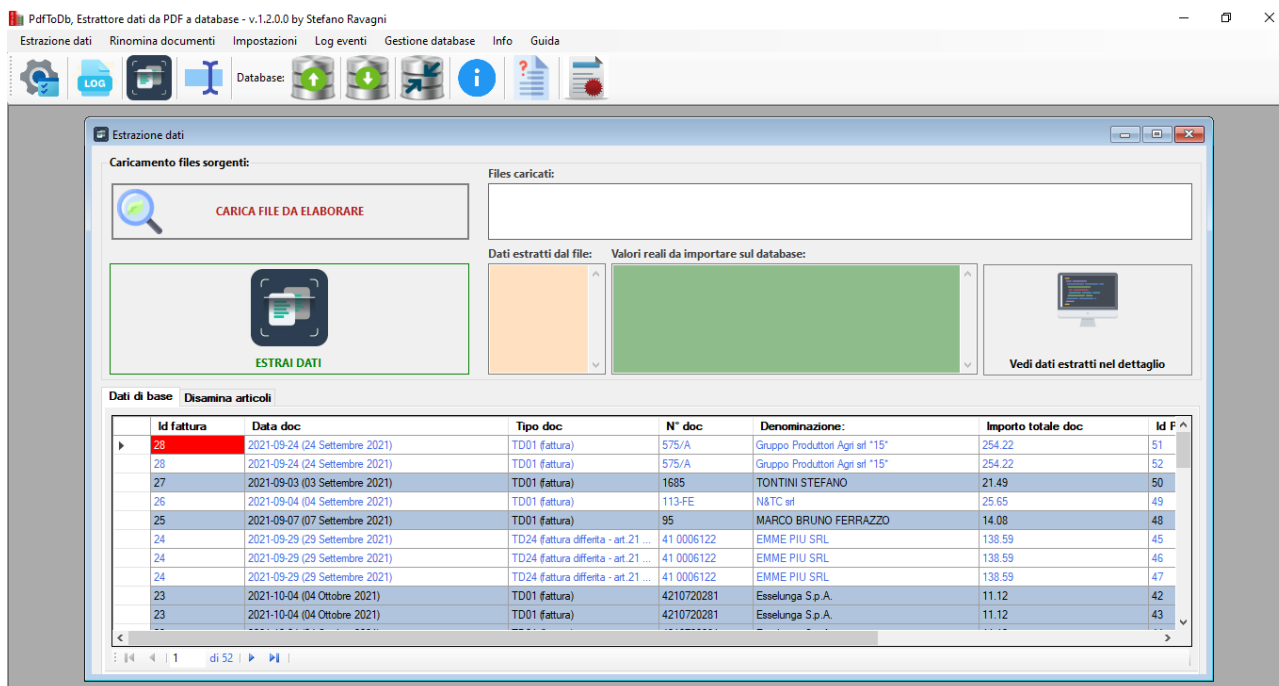


# PdfToDb, Estrazione di dati da PDF e salvataggio su database

Sviluppato da Stefano Ravagni

V 1.3.1.0 – Marzo 2023



- Cos'è PdfToDb..... 3
- Come funziona PdfToDb..... 3
- Configurazione di PdfToDb ..... 3
  - Impostazioni database: ..... 3
    - Uso del percorso di default per il database: ..... 4
    - Password del database:..... 4
    - Abilita auto-backup: ..... 4
    - Prefisso per il nome del file di backup da creare: ..... 4
    - Percorso alternativo per backup e salvataggio dei documenti:..... 5
    - Password per il backup: ..... 5
  - Impostazioni campi database estrazione dei dati di base: ..... 6
  - Impostazioni campi database estrazione disamina prodotti: ..... 8
    - Campo speciale: ..... 9
  - Impostazioni per rinomina documenti: ..... 10
    - Percorso di salvataggio dei documenti rinominati:..... 13
  - Impostazioni generali: ..... 14
- Estrazione dati..... 15
  - La griglia dati ..... 17
  - Il menu contestuale delle griglie dati ..... 18

# PdfToDb, Estrazione di dati da PDF e salvataggio su database

Sviluppato da Stefano Ravagni

V 1.3.1.0 – Marzo 2023

Visualizza dati.....	18
Modifica dati .....	19
Elimina documento: .....	22
Ricarica dati sulla griglia:.....	22
Inserisci campo speciale:.....	22
Svuota tabella:.....	23
Rinomina documenti.....	24
Log eventi .....	25
Importazione database .....	25
Esportazione database .....	26
Caratteristiche tecniche .....	26
Licenza d'uso (EULA) .....	27
Ringraziamenti .....	27

# PdfToDb, Estrazione di dati da PDF e salvataggio su database

Sviluppato da Stefano Ravagni

V 1.3.1.0 – Marzo 2023

## Cos'è PdfToDb

**PdfToDb** è un software per l'estrazione di dati a scelta dell'utente da files PDF e successivo salvataggio su un database (attualmente MsAccess).

I dati estratti dai files sorgente in formato PDF saranno scegli dall'utente attraverso la configurazione di alcuni campi obbligatori ed altri 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 in base alle stringhe di testo indicate dall'utente; 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 dall'utente nelle impostazioni.

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

## Configurazione di PdfToDb

Le impostazioni di PdfToDb sono raggiungibili dalla toolbar o dal menu relativi alle impostazioni. La form delle impostazioni è attualmente suddivisa in 4 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.

# PdfToDb, Estrazione di dati da PDF e salvataggio su database

Sviluppato da Stefano Ravagni

V 1.3.1.0 – Marzo 2023

The screenshot shows the 'Impostazioni' window with the 'Impostazioni database' tab selected. The window title is 'Impostazioni'. The tab bar includes 'Impostazioni database', 'Impostazioni campi database estrazione dati di base', 'Impostazioni campi database estrazione disamina prodotti', and 'Impostazioni generali'. The 'Database di destinazione dati- MsAccess' section has a checked checkbox for 'Usa percorso di default per il database' and a 'Password del database' field. Below this is a note: 'Deselezionando questo checkbox sarà possibile specificare con quale database lavorare e su quale percorso anziché utilizzare le impostazioni di default'. The 'Percorso alternativo database' field contains 'C:\Users\Stefano\Desktop\' and the 'Nome del file database' field contains 'PdfToDb.mdb'. The 'Backup database:' section has a green warning: 'Attenzione ! Il backup è attualmente previsto per MsAccess ed ha come destinazione di default la cartella "BackupPdfToDb" salvo diversa indicazione dell'Destinatario da apportare tramite le sottostanti caselle di controllo.' It includes an 'Abilita Auto Backup' checkbox (unchecked), a 'Prefisso per il nome del file di backup da creare' field with 'POSTAZIONE\_1', a 'Percorso alternativo per backup e salvataggio dei documenti' field with a 'Cerca percorso' button, and a 'Password per il Backup' field with masked characters. A 'Salva configurazione' button is at the bottom right.

## **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 database:**

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

# PdfToDb, Estrazione di dati da PDF e salvataggio su database

Sviluppato da Stefano Ravagni

V 1.3.1.0 – Marzo 2023

## ***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.

# PdfToDb, Estrazione di dati da PDF e salvataggio su database

Sviluppato da Stefano Ravagni

V 1.3.1.0 – Marzo 2023

## Impostazioni campi database estrazione dei dati di base:

**Impostazioni**

Impostazioni database | **Impostazioni campi database estrazione dati di base** | Impostazioni campi database estrazione disamina prodotti | Impostazioni generali

**Campi da estrarre dal file sorgente:**

Inserire i nomi dei campi del file sorgente dai quali estrarre i dati: è necessario inserire il campo esatto antecedente il valore da estrarre comprensivo del carattere di separazione. Esempio: per estrarre il valore del campo "Denominazione" inserire "Denominazione:". L'elemento separatore, in questo caso i due punti, dipende da come è formattato il documento.

Tutti i campi **OBBLIGATORI** devono essere valorizzati, per i sottocampi **AGGIUNTIVI** inserirne un minimo di due partendo dal primo e lasciare gli altri vuoti se non necessari.

Importante !!! In caso di stringhe identiche da ricercare all'interno del documento sorgente sarà presa in considerazione solamente la prima per i campi obbligatori, saranno invece usate tutte quante per i campi opzionali.

I sottocampi **AGGIUNTIVI** sono da intendersi come possibili valori **MULTIPLI** legati ai campi obbligatori, la dove ci siano valori ripetuti da estrarre (es: IVA APPLICATA e **IMPONIBILE** per ogni prodotto in un documento di fatturazione).

	Nome campi obbligatori	Etichetta visibile	Spazio	Sottocampi aggiuntivi	Etichetta visibile	Spazio
Campo 1 (Tipo STRINGA):	Data documento:	Data doc	280	Aliquota IVA (%):	IVA %	70
Campo 2 (Tipo STRINGA):	Tipologia documento:	Tipo doc	160	Totale imponibile/importo:	Importo imp	100
Campo 3 (Tipo STRINGA):	Numero documento:	N° doc	100			0
Campo 4 (Tipo STRINGA):	Denominazione:	Denominazione:	240			0
Campo 5 (Tipo STRINGA):	Importo totale documento:	Importo totale doc	180			0

Salva configurazione

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 **la 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 succedrebbe se venisse trovata la seguente riga....

# PdfToDb, Estrazione di dati da PDF e salvataggio su database

Sviluppato da Stefano Ravagni

V 1.3.1.0 – Marzo 2023

**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 una 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 impostazione 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.

# PdfToDb, Estrazione di dati da PDF e salvataggio su database

Sviluppato da Stefano Ravagni

V 1.3.1.0 – Marzo 2023

## Impostazioni campi database estrazione disamina prodotti:

**Impostazioni**

Impostazioni database | Impostazioni campi database estrazione dati di base | **Impostazioni campi database estrazione disamina prodotti** | Impostazioni generali

**Campi da estrarre dal file sorgente:**

Inserire i nomi dei campi del file sorgente dai quali estrarre i dati: è necessario inserire il campo esatto antecedente il valore da estrarre comprensivo del carattere di separazione. Esempio: per estrarre il valore del campo "Denominazione" inserire "Denominazione:". L'elemento separatore, in questo caso i due punti, dipende da come è formattato il documento.

Tutti i campi **OBBLIGATORI** devono essere valorizzati, per i sottocampi **AGGIUNTIVI** inserirne un minimo di due partendo dal primo e lasciare gli altri vuoti se non necessari.

Importante !!! In caso di stringhe identiche da ricercare all'interno del documento sorgente sarà presa in considerazione solamente la prima per i campi obbligatori, saranno invece usate tutte quante per i campi opzionali.

I sottocampi **AGGIUNTIVI** sono da intendersi come possibili valori **MULTIPLI** legati ai campi obbligatori, la dove ci siano valori ripetuti da estrarre (es: IVA APPLICATA e IMPONIBILE per ogni prodotto in un documento di fatturazione).

Stringa da usare come blocco di partenza:  Stringa di partenza inizio estrazione prodotti:

	Nome campi obbligatori	Etichetta visibile	Spazio	Sottocampi aggiuntivi	Etichetta visibile	Spazio
Campo 1 (Tipo STRINGA):	<input type="text" value="Data documento:"/>	<input type="text" value="Data doc"/>	<input type="text" value="160"/>	<input type="text" value="Descrizione bene/servizio:"/>	<input type="text" value="Prodotto/Sconto"/>	<input type="text" value="280"/>
Campo 2 (Tipo STRINGA):	<input type="text" value="Tipologia documento:"/>	<input type="text" value="Tipo doc"/>	<input type="text" value="130"/>	<input type="text" value="Quantità:"/>	<input type="text" value="Quantità"/>	<input type="text" value="80"/>
Campo 3 (Tipo STRINGA):	<input type="text" value="Numero documento:"/>	<input type="text" value="N° doc"/>	<input type="text" value="100"/>	<input type="text" value="Valore totale:"/>	<input type="text" value="Val totale"/>	<input type="text" value="100"/>
Campo 4 (Tipo STRINGA):	<input type="text" value="Denominazione:"/>	<input type="text" value="Denominazione:"/>	<input type="text" value="200"/>	<input type="text" value="IVA (%):"/>	<input style="background-color: #cccccc;" type="text" value="IVA (%)"/>	<input type="text" value="80"/>
Campo 5 (Tipo STRINGA):	<input type="text" value="Importo totale documento:"/>	<input type="text" value="Importo totale doc"/>	<input type="text" value="140"/>	<input style="background-color: #cccccc;" type="text" value=""/>	<input style="background-color: #cccccc;" type="text" value=""/>	<input type="text" value="0"/>

**Campo speciale**

**Salva configurazione**

La disamina degli articoli serve sostanzialmente per analizzare una serie di campi relativi ad articoli e/o prodotti per i quali è stato applicato uno sconto; la disamina riesce ad estrarre sia i dati di interesse generale tramite i campi fissi come già visto per la analisi di base, che i campi aggiuntivi che presentano i dati relativi agli sconti applicati.

Come per l'estrazione dei dati di base, anche questa sezione delle impostazioni consente all'utente di impostare un sistema di ricerca ed estrazione dei dati partendo dai file sorgente PDF ed usa nello stesso modo sia campi fissi obbligatori che campi opzionali aggiuntivi.

A differenza della estrazione dei dati di base però, l'estrazione per disamina articoli consente di impostare un blocco di partenza sul file originale, scavalcando così gran parte del testo non utile che allungherebbe i tempi di elaborazione.

Consente inoltre di specificare una stringa di inizio per l'estrazione dei prodotti, che possono essere anche moltissimi e su pezzi di testo doppi, come nel caso di un prodotto descritto dapprima nelle specifiche fisse e successivamente nelle specifiche degli sconti applicati.



# PdfToDb, Estrazione di dati da PDF e salvataggio su database

Sviluppato da Stefano Ravagni

V 1.3.1.0 – Marzo 2023

Al di là dei due campi STRINGA DA USARE COME BLOCCO DI PARTENZA e STRINGA DI PARTENZA INIZIO ESTRAZIONE PRODOTTI valgono le stesse indicazioni già fornite per la analisi dei dati di base.

Si fa inoltre presente che la disamina dei prodotti può essere disabilitata dalla tab delle impostazioni generali se non fosse utile all'utente.

## **Campo speciale:**

Il campo speciale è un campo adattabile agli interessi dell'utente ed esula dai valori estratti dai documenti di interesse... potrebbe essere un barcode, un numero, o qualsiasi altro dato che l'utente voglia legare ai prodotti della disamina per proprio tornaconto...

Come per gli altri campi è possibile assegnargli una etichetta, in modo da poterla cambiare in futuro con altri tipi di dati più utili.

# PdfToDb, Estrazione di dati da PDF e salvataggio su database

Sviluppato da Stefano Ravagni

V 1.3.1.0 – Marzo 2023

## Impostazioni per rinomina documenti:

A partire dalla versione 1.3.0.0 è stata introdotta la funzionalità per la rinomina dei documenti, che consente all'utente di creare un sistema di rinomina automatica in base al contenuto di ogni file, ovviamente totalmente configurabile.

The screenshot shows the 'Impostazioni' window with the 'Rinomina documenti' tab selected. The interface includes a search bar for documents, a list of extracted lines, and several configuration fields:

- Campi da estrarre dal file sorgente:** A section with a blue 'I' icon and explanatory text about primary and sequential data extraction.
- Cerca documento per selezionare i dati da estrarre:** A search input field with the text 'Nessun documento caricato'.
- Righe estratte dal documento caricato per estrazione:** A dropdown menu.
- Dati primari:** Fields for 'Stringa da cercare per inizio estrazione' (Denominazione), 'Occorrenza da usare' (2), 'Stringhe da escludere' (Denominazione), 'Estrai testo fino alla stringa' (Indirizzo), and 'Ordine del blocco' (Ultimo).
- Dati in sequenza:** Fields for 'Stringa intestazione colonna da cercare' (tipologia documento), 'Carattere divisione testo' (SPAZIO), 'Stringhe da escludere' (differita di cui all'art), and 'Blocchi da estrarre' (1, 2, 3, 4, 5).
- Trasforma i dati di tipo DATA in formato:** A dropdown menu set to 'Formato anglosassone'.
- Ordine x rinomina:** A row of dropdown menus for blocks 2, 3, 4, 1, and 5.
- Carattere di separazione tra i campi nel nome finale:** A field containing an underscore.
- Pagina del documento da analizzare:** A field containing the number 1.
- Analizza e testa il risultato della rinomina:** A green button.
- Risultato della combinazione dei campi in linea teorica:** A green text label.
- Percorso di salvataggio dei documenti rinominati:** A search input field with the path 'C:\Users\User\Desktop\vinomine'.
- Salva configurazione:** A blue floppy disk icon button.

Tramite il configuratore è possibile indicare tutti i parametri necessari a spezzettare il nome del file finale in base ai contenuti del documento; il tutto è utilizzabile secondo i seguenti step.

1. Caricare un documento in formato PDF tramite il bottone **CERCA DOCUMENTO PER SELEZIONARE I DATI DA ESTRARRE**. Una volta caricato almeno un documento la combobox dal titolo Righe estratte dal documento caricato per estrazione si popolerà di tutte le righe di cui è fatto il documento, consentendo all'utente di comprendere quali righe si hanno a disposizione e impostare il programma nel modo corretto ai propri desideri.

# PdfToDb, Estrazione di dati da PDF e salvataggio su database

Sviluppato da Stefano Ravagni

V 1.3.1.0 – Marzo 2023

**Campi da estrarre dal file sorgente:**

Caricare un documento e selezionare per i DATI PRIMARI e per i DATI IN SEQUENZA le righe da associare tra quelle presenti nella combobox RIGHE ESTRATTE DAL DOCUMENTO CARICATO PER ESTRAZIONE usando i bottoni ASSOCIA LA RIGA SELEZIONATA PER ESTRAZIONE.  
Per i dati primari inserire una eventuale stringa che funga da delimitatore... in presenza di tale stringa viene estratto il testo della riga associata fino a quando non si incontra la stringa aggiuntiva inserita.  
Per i dati in sequenza inserire un carattere di divisione del testo in modo da suddividere la riga scelta in più blocchi: inserire per i blocchi desiderati (da 1 a 4) il numero del blocco e la relativa ordinalità del blocco in modo da costruire un nome finale del file secondo le proprie esigenze.

**Righe estratte dal documento caricato per estrazione:**

0	Cedente/prestatore (fornitore) Cessionario/committente (cliente)
1	Identificativo fiscale ai fini IVA: IT01191091006 Identificativo fiscale ai fini IVA: IT02515790596
2	Codice fiscale: 03373860588 Denominazione: N&TC S.R.L.
3	Denominazione: FA.LU.CIOLI S.R.L. Indirizzo: V. GIACOMO MATTEOTTI 126
4	Regime fiscale: RF01 ordinario Comune: APRILIA Provincia: LT
5	Indirizzo: VIA DELL'ELETTRONICA, 16 Cap: 04011 Nazione: IT
6	Comune: ROMA Provincia: RM
7	Cap: 00144 Nazione: IT
8	Telefono: 069330125
9	Email: info@cioli.com
10	Tipologia documento Art. 73 Numero documento Data documento Codice destinatario
11	TD01 fattura 023890-0C0 03-11-2022 USAL8PV
12	Causale
13	FATT. VENDITA MERCEASSOLVE GLI OBBLIGHI DI CUI ALL'ART. 62 COMMA 1 DEL DL24/01/2012 N 1 CONVERTITO CON MODIFICAZIONI
14	DALLA LEGGE DEL24/03//2012 N 27
15	Sconto o
16	Cod. articolo Descrizione Quantità Prezzo unitario UM %IVA Prezzo totale
17	magg.
18	30 (Codice interno)
19	GUANCIALE AL PEPE LOTTO:031122 5,50 6,50 KG 10,00 35,75
20	30 (AswArtFor)
21	Tipo dato: AswConCont
22	Rif. testo: 500500
23	Tipo dato: AswCenCost
24	Rif. testo: 0309001
25	32 (Codice interno)
26	COPPIETTE TIPICHE LOTTO:031122 1,00 18,00 KG 10,00 18,00
27	32 (AswArtFor)
28	Tipo dato: AswConCont
29	Rif. testo: 500500

2. Impostare come vanno utilizzati i **dati primari**, intendendo con prima quelli di maggiore rilevanza (a scelta dell'utente);
  - a. immettere la **stringa da cercare per inizio estrazione** per i dati fondamentali
  - b. immettere il **numero di occorrenza da utilizzare...** questo parametro è utile qualora la stringa da cercare fosse presente più volte, nel quale caso si va a specificare quale è quella necessaria, escludendo le altre... nell'esempio viene ricercata la stringa denominazione:, specificando che vogliamo la seconda occorrenza della stessa... come è possibile notare dalla casella combinata la stringa in questione è presente sia alla riga numero 2 che sulla numero 3 e specificando che vogliamo la seconda sarà proprio la terza riga ad essere utilizzata come dati primari.
  - c. Immettere le **stringhe da escludere**, che vanno interpretate come le stringhe da tagliare via una volta che si è capito quale sia la riga da analizzare... in questo caso si va a dire al programma che vogliamo si partire dalla stringa denominazione: ma anche che tale stringa alla fine andrà esclusa dal risultato finale. Puoi avere più di una stringa a patto di suddividerle con il carattere PIPE, ovvero il simbolo |, senza spazi; ogni stringa verrà interpretata, cercata ed esclusa dal risultato finale.
  - d. Immettere la **stringa limite per la riga da analizzare**, ovvero la riga di termine estrazione partendo da quella inizialmente impostata; nell'esempio si parte dalla stringa denominazione: e si indica di voler estrarre fino a che si incontra la stringa indirizzo: ... infatti alla riga 3, quella che dovrà come già spiegato essere analizzata, è presente la stringa intera Denominazione: FA.LU.CIOLI S.R.L. Indirizzo: V. GIACOMO MATTEOTTI 126 ... bene... finora abbiamo detto al programma di partire da Denominazione: ed estrarre tutto il testo che si

# PdfToDb, Estrazione di dati da PDF e salvataggio su database

Sviluppato da Stefano Ravagni

V 1.3.1.0 – Marzo 2023

- incontra fino a che non si trova la stringa Indirizzo:, per cui il risultato della estrazione sarà Denominazione: FA.LU.CIOLI S.R.L., salvo poi escludere la stringa Denominazione:, lasciando come risultato ultimo il nome che ci interessa, ovvero FA.LU.CIOLI S.R.L.
- e. I dati primari nella rinomina del file saranno potranno trovarsi in prima posizione o in ultima posizione rispetto alla nomenclatura del file rinominato finale, in base alle impostazioni dell'utente; per scegliere questa impostazione basta selezionare dalla combo box se l'ordine del box prima deve stare in prima posizione o in ultima.
3. Impostare l'interpretazione dei **dati in sequenza**, intendendo i dati di dettaglio del documento che potranno essere scomposti e disposti a piacere nel nome finale del file rinominato.
- a. indicare la **stringa di intestazione colonna da cercare**, ovvero la stringa presente nella riga che fa intestazione di colonna per i dati che seguiranno... è infatti attualmente previsto che tali dati siano disposti in colonna anziché per riga. Nel nostro esempio si specifica la stringa **Tipologia documento** presente nella riga 10 in quanto, controllando il format dei documenti da analizzare, è la prima voce nonché la prima colonna della tabella che porta i dati di dettaglio, che si trovano sulla riga 11... specificando cosa cercare per capire dove inizia la tabella sappiamo anche che dalla riga successiva troveremo i dati sequenziali che ci interessano.
- b. Impostare il **carattere di divisione del testo**; abbiamo capito che i dati si trovano nella riga 11 e corrispondono ai seguenti dati: TD01 fattura 023890-0C0 03-11-2022 USAL8PV... osservando questa stringa dobbiamo trovare qualcosa di univoco che suddivida i dati in modo da poter spezzare tale stringa in tante sotto parti... può essere un puntino, una virgola, un altro carattere o un insieme di caratteri... in questo caso i dati sono separati da uno spazio, per cui nel campo metterò uno spazio premendo una volta sulla barra spaziatrice della tastiera... essendo un carattere invisibile, viene trasformato nella dicitura SPAZIO per una migliore comprensione per l'utente.
- c. Immettere le **stringhe da escludere**, che vanno interpretate come le stringhe da tagliare via una volta dal file rinominato finale... Utile per eliminare tutte quelle stringhe che a volte danno fastidio, qualunque siano. Puoi avere più di una stringa da escludere **a patto di suddividerle con il carattere PIPE**, ovvero il simbolo |, senza spazi; ogni stringa verrà interpretata, cercata ed esclusa dal risultato finale.
- d. Impostare la sequenza dei blocchi da analizzare; la stringa spezzettata grazie al carattere di divisione del testo, può essere rimontata indicando la sequenza in cui ogni blocco andrà disposto... grazie a questo sistema è possibile fare delle rinomine puntuali e personalizzate secondo quelle che sono le esigenze dell'utente. Per fare qualche esempio esplicativo, la stringa TD01 fattura 023890-0C0 03-11-2022 USAL8PV può potrà essere usata come fattura 03-11-2022

# PdfToDb, Estrazione di dati da PDF e salvataggio su database

Sviluppato da Stefano Ravagni

V 1.3.1.0 – Marzo 2023

TD01 USAL8PV 023890-0C0 oppure come USAL8PV TD01 03-11-2022 fattura 023890-0C0 o come altre combinazioni. Qualora un blocco non fosse necessario o gradito all'utente, **sarà sufficiente immettere il valore 0 al blocco stesso**, così che il software non lo prenderà in considerazione.

4. Impostare l'utilizzo della **trasformazione dei dati di tipo DATA/TEMPO**: questa impostazione consente di indicare al programma di trasformare le stringhe che rappresentano delle date valide in un formato italiano o anglosassone. I valori possibili sono attualmente solo due e devono essere scelti dalla combobox presente.
5. Impostare il **carattere di separazione tra i campi nel nome finale**: è il carattere che verrà automaticamente aggiunto per separare i blocchi dei dati nel nome finale del file, se desiderato dall'utente...altrimenti è possibile lasciare il campo vuoto. Può essere un carattere qualsiasi o un insieme di caratteri.
6. Impostare il **numero della pagina del documento da analizzare**: dato che i file pdf utilizzati potrebbero essere anche molto grandi, si indica al programma quale pagina analizzare per elaborare tutto quanto è stato detto.

Il bottone **analizza e testa il risultato della rinomina** mostra una anteprima del risultato in base ai parametri inseriti dall'utente, in modo che possa capire se vanno bene o modificarli aggiustando il tiro fino al raggiungimento del risultato desiderato.

## Percorso di salvataggio dei documenti rinominati:

È possibile indicare un percorso, locale o di rete, sul quale i file rinominati verranno salvati in automatico. Da notare bene che **un percorso deve esistere sempre**.

Quando tutto sembra funziona come desiderato **ricordarsi di salvare le impostazioni**.

# PdfToDb, Estrazione di dati da PDF e salvataggio su database

Sviluppato da Stefano Ravagni

V 1.3.1.0 – Marzo 2023

## Impostazioni generali:



Tra le impostazioni generali sono presenti i seguenti campi:

### Separatore di campi documento PDF originale:

Necessario per il delimitatore dei campi ma attualmente è **non utilizzato**.

### Usa sistema per analisi e disamina prodotti:

serve per abilitare o disabilitare la seconda sorgente dati e relativa griglia per una analisi più approfondita dei documenti, chiamata DISAMINA PRODOTTI... per “prodotti” si intende qualsiasi dato di interesse per l’utente finale.

La voce è attivata per default.

### Usa finestre massimizzate di default:

serve per indicare al programma di utilizzare la massimizzazione dimensionale delle finestre figlie all’interno del software, là dove questo sia applicabile.

Attualmente solo la finestra principale di estrazioni dati e quella della visualizzazione dei dati sono ridimensionabili.

### Lancia estrazione dati all’avvio del programma:

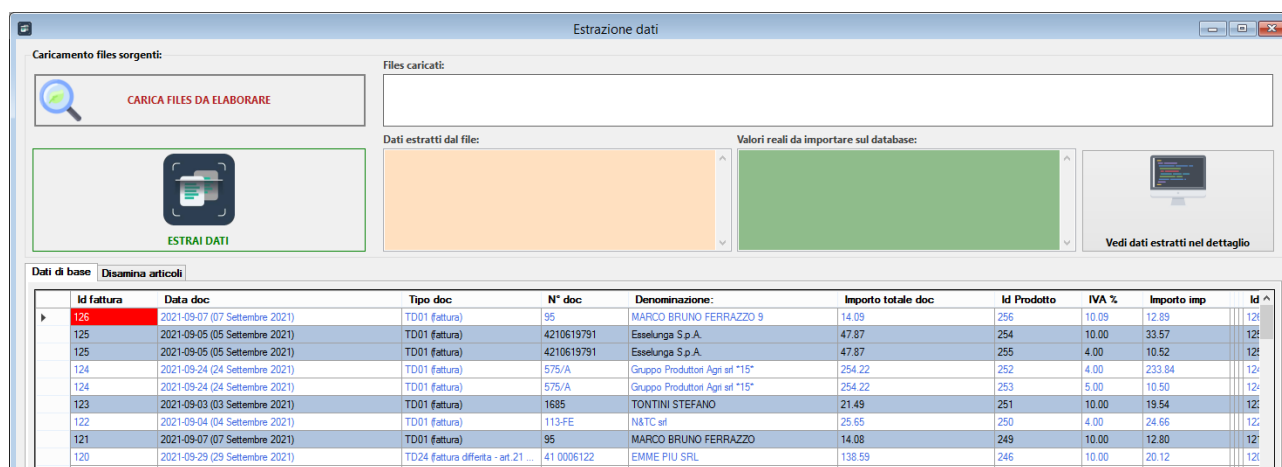
# PdfToDb, Estrazione di dati da PDF e salvataggio su database

Sviluppato da Stefano Ravagni

V 1.3.1.0 – Marzo 2023

Indica al programma se deve aprire automaticamente o meno la finestra principale relativa alla estrazione dati al momento dell'apertura del software. Dato che la schermata può presentare un numero elevato di dati si consiglia di lasciare disabilitata questa funzione in modo da rendere il software più veloce in fase di avvio; la funzionalità di estrazione dei dati sarà richiamabile manualmente dall'utente con un semplice click della toolbar.

## Estrazione dati



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

Se l'utente ha abilitato l'uso della disamina prodotti, potrà eseguire due tipi di estrazione diverse, che comunque seguono i seguenti step logici:

1. Cliccare sul bottone **CARICA FILE DA ELABORARE**: questo permetterà il caricamento di 1 o N file 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. I dati saranno estratti in relazione alle impostazioni configurate dall'utente, che sceglie in apposite schermate delle impostazioni quali stringhe di testo ricercare e salvare, come visualizzarle e come interpretarle.

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

# PdfToDb, Estrazione di dati da PDF e salvataggio su database

Sviluppato da Stefano Ravagni

V 1.3.1.0 – Marzo 2023

The screenshot shows the 'Estrazione dati' window. It has a top bar with the title 'Estrazione dati'. Below it, there are two main sections: 'Caricamento files sorgenti' and 'Files caricati'. The 'Caricamento files sorgenti' section has a search icon and a button labeled 'CARICA FILE DA ELABORARE'. The 'Files caricati' section shows a file named 'TEST\_IT01378570350\_VrM9z.PDF'. Below these are two columns: 'Dati estratti dal file' and 'Valori reali da importare sul database'. The 'Dati estratti dal file' column shows the file path and document details like 'FATTURA ELETTRONICA Versione FPR12'. The 'Valori reali da importare sul database' column shows the same file path and document details. To the right of these columns is a button labeled 'Vedi dati estratti nel dettaglio'. At the bottom, there is a table titled 'Dati di base' with a sub-header 'Disamina articoli'. The table has columns: 'Id fattura', 'Data doc', 'Tipo doc', 'N° doc', 'Denominazione:', 'Importo totale doc', 'Id Prodotto', 'IVA %', 'Importo imp', and 'Id'. The table contains several rows of data, with the first row highlighted in red.

Id fattura	Data doc	Tipo doc	N° doc	Denominazione:	Importo totale doc	Id Prodotto	IVA %	Importo imp	Id
128	2021-09-05 (05 Settembre 2021)	TD01 (fattura)	4210619791	Esselunga S.p.A.	47.87	257	10.00	33.57	128
128	2021-09-05 (05 Settembre 2021)	TD01 (fattura)	4210619791	Esselunga S.p.A.	47.87	258	4.00	10.52	128
126	2021-09-07 (07 Settembre 2021)	TD01 (fattura)	95	MARCO BRUNO FERRAZZO 9	14.09	256	10.09	12.89	126
125	2021-09-05 (05 Settembre 2021)	TD01 (fattura)	4210619791	Esselunga S.p.A.	47.87	254	10.00	33.57	125
125	2021-09-05 (05 Settembre 2021)	TD01 (fattura)	4210619791	Esselunga S.p.A.	47.87	255	4.00	10.52	125
124	2021-09-24 (24 Settembre 2021)	TD01 (fattura)	575/A	Gruppo Produttori Agri srl "15"	254.22	252	4.00	233.84	124
124	2021-09-24 (24 Settembre 2021)	TD01 (fattura)	575/A	Gruppo Produttori Agri srl "15"	254.22	253	5.00	10.60	124

Sul box **DATI ESTRATTI DAL FILE** compariranno tutte le righe di ogni file 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.

The screenshot shows a detailed view of the data extraction process for a specific file. At the top, there is a PDF icon and the file name 'IT01378570350\_VrM9z.PDF'. Below this, there are two main sections: 'Dati estratti dal file' and 'Valori reali da importare sul database'. The 'Dati estratti dal file' section shows the file path and document details like 'FATTURA ELETTRONICA Versione FPR12'. The 'Valori reali da importare sul database' section shows the same file path and document details. To the right of these sections is a button labeled 'Vedi dati estratti nel dettaglio'.

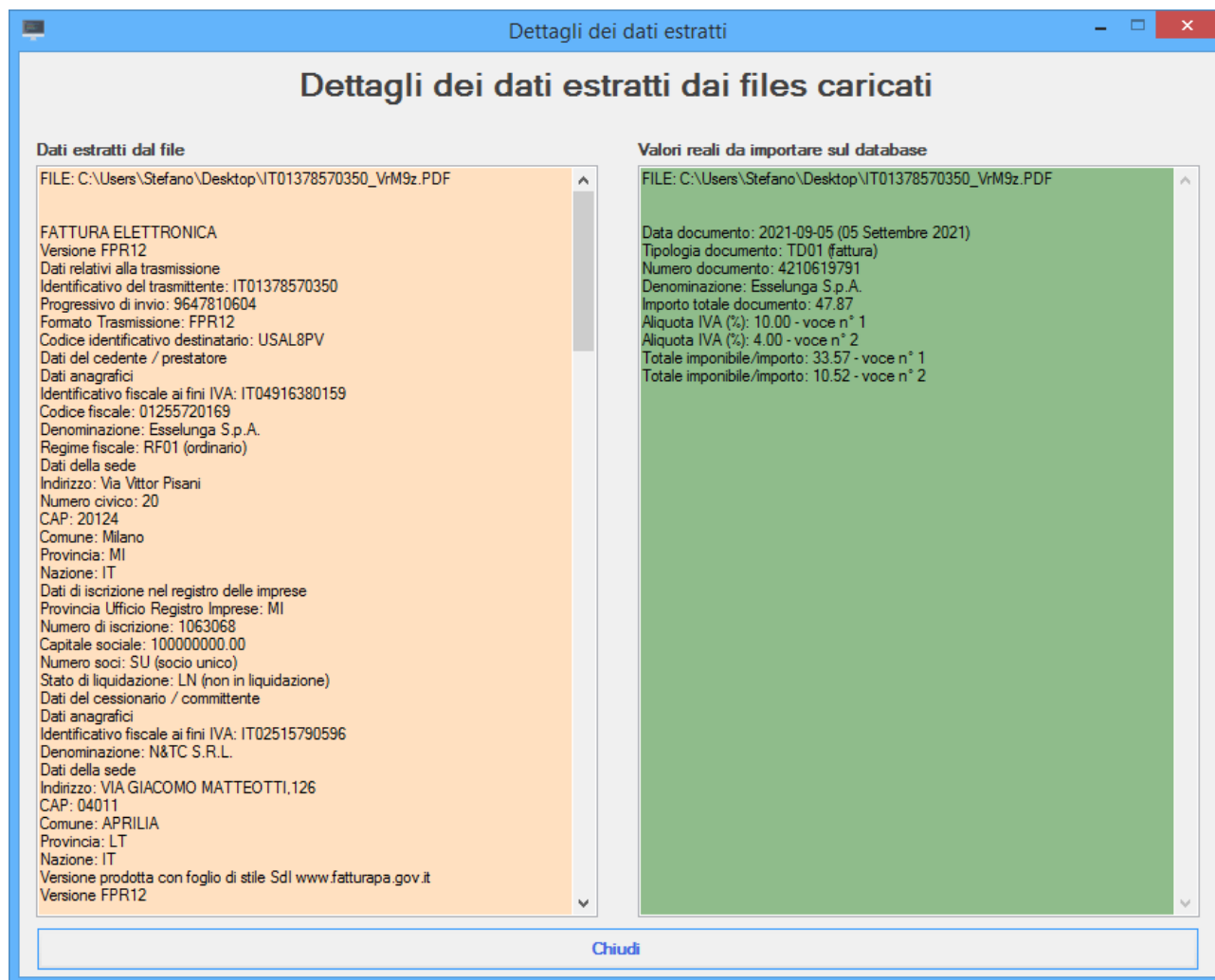
È possibile vedere questi dati in dettaglio o scorrendo i singoli box o cliccando sul bottone **VEDI DATI ESTRATTI NEL DETTAGLIO**.



# PdfToDb, Estrazione di dati da PDF e salvataggio su database

Sviluppato da Stefano Ravagni

V 1.3.1.0 – Marzo 2023



## 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.

# PdfToDb, Estrazione di dati da PDF e salvataggio su database

Sviluppato da Stefano Ravagni

V 1.3.1.0 – Marzo 2023

	Id fattura	Data doc	Tipo doc	N° doc	Denominazione:
▶	128	2021-09-05 (05 Settembre 2021)	TD01 (fattura)	4210619791	Esselunga S.p.A.
	128	2021-09-05 (05 Settembre 2021)	TD01 (fattura)	4210619791	Esselunga S.p.A.
	126	2021-09-07 (07 Settembre 2021)	TD01 (fattura)	95	MARCO BRUNO FERRAZZO 9
	125	2021-09-05 (05 Settembre 2021)	TD01 (fattura)	4210619791	Esselunga S.p.A.
	125	2021-09-05 (05 Settembre 2021)	TD01 (fattura)	4210619791	Esselunga S.p.A.
	124	2021-09-24 (24 Settembre 2021)	TD01 (fattura)	575/A	Gruppo Produttori Agri srl "15"
	124	2021-09-24 (24 Settembre 2021)	TD01 (fattura)	575/A	Gruppo Produttori Agri srl "15"
	123	2021-09-03 (03 Settembre 2021)	TD01 (fattura)	1685	TONTINI STEFFANO

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

Si noti il raggruppamento dei colori delle righe che variano in base all'id della fattura (1° colonna), in modo da raggruppare le acquisizioni dei dati relativi alla medesima fattura e differenziarle in modo alternato dalle precedenti e dalle successive. Le righe selezionate vengono invece evidenziate con una colorazione rossa ben visibile.

## Il menu contestuale delle griglie dati

Cliccando con il tasto destro su una delle due griglie (Dati di base o Disamina articoli) compare un menu contestuale con molte funzioni utili ed importanti.



## Visualizza dati

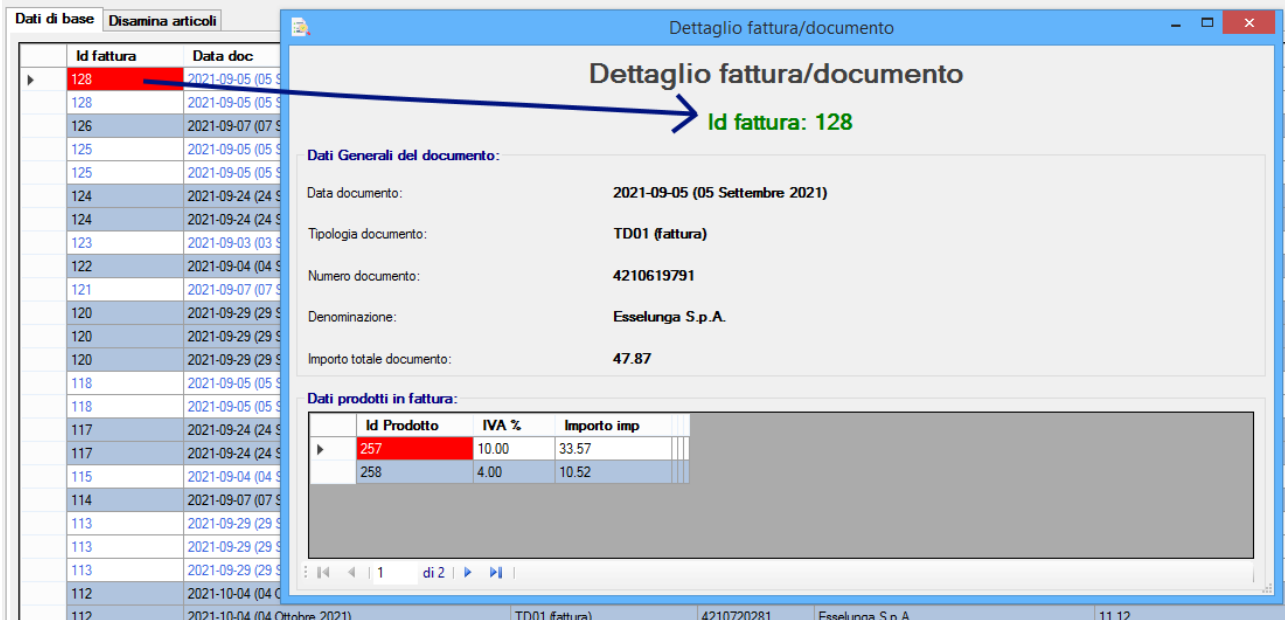
Con visualizza dati l'utente può andare a vedere una form con all'interno il dettaglio dei dati relativi alla fattura selezionata... ecco due screenshot per i dati di base e per la disamina articoli

# PdfToDb, Estrazione di dati da PDF e salvataggio su database

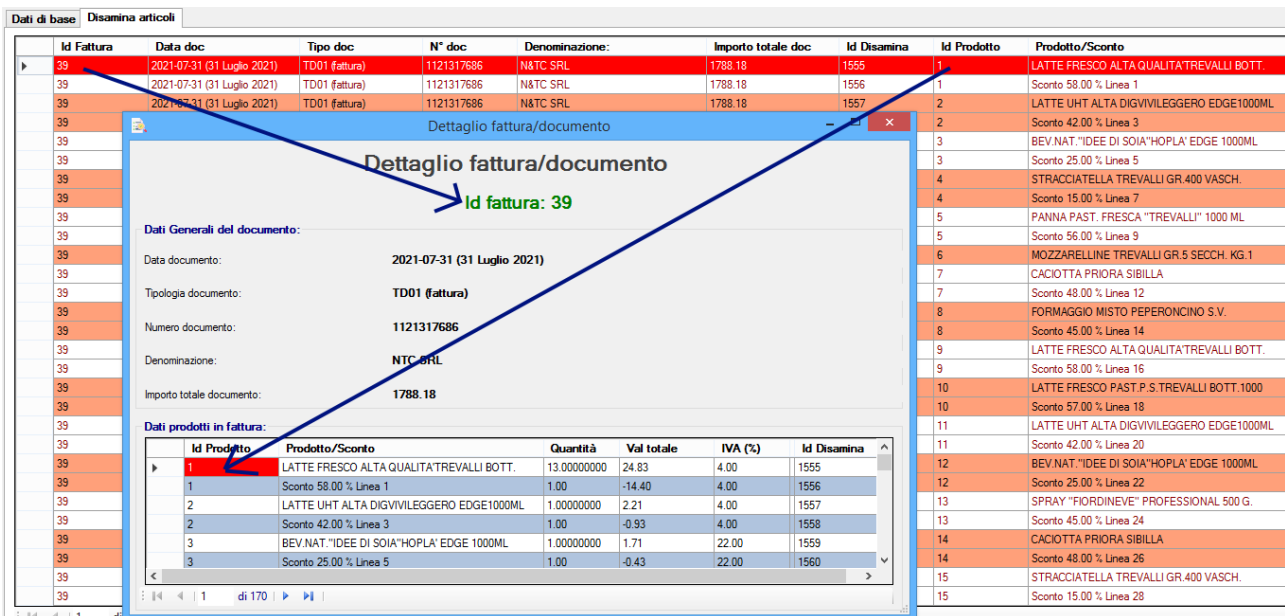
Sviluppato da Stefano Ravagni

V 1.3.1.0 – Marzo 2023

Dati di base:



Disamina articoli:



La form della visualizzazione dati è ridimensionabile dall'utente.

## Modifica dati

La modifica dati consente all'utente di andare a modificare i dati estratti e salvatati; visto che l'operazione è automatica e che c'è la possibilità che alcuni dati siano stati estratti in modo errato o parziale, viene data all'utente la possibilità di modificarli.

# PdfToDb, Estrazione di dati da PDF e salvataggio su database

Sviluppato da Stefano Ravagni

V 1.3.1.0 – Marzo 2023

Per usare questa funzione basta usare il menu contestuale e selezionare la voce modifica dati, sia dai dati di base che da quelli di disamina.

**Modifica dati fattura**

**Modifica dati acquisiti**

**Id fattura: 128**

**Dati generali del documento:**

Data documento: 2021-09-05 (05 Settembre 2021)

Tipologia documento: TD01 (fattura)

Numero documento: 4210619791

Denominazione: Esselunga S.p.A.

Importo totale documento: 47.87

**Dati prodotti in fattura:**

ID	IVA %	Importo imp			
257	10.00	33.57			
258	4.00	10.52			
ID					
ID					
ID					

Annulla modifiche | Salva modifiche

È possibile modificare sia i dati generali del documento che quelli legati ai prodotti in fattura; una volta apportate le modifiche queste potranno essere consolidate con la pressione del bottone **SALVA MODIFICHE**.

Sulla modifica richiamata dalla griglia della disamina, viene applicato un filtro che lega i dati alla fattura selezionata; inoltre viene eseguito un filtro sul n° della disamina e sull'id del prodotto, mostrando quindi solo i prodotti selezionati all'interno di una fattura che potrebbe contenerne moltissimi... per rendere più agevole la modifica e la visualizzazione vengono fatti modificare solo i dati selezionati dall'utente.

# PdfToDb, Estrazione di dati da PDF e salvataggio su database

Sviluppato da Stefano Ravagni

V 1.3.1.0 – Marzo 2023

Dati di base Disamina articoli

Id Fattura	Data doc	Tipo doc	N° doc	Denominazione:	Importo totale doc	Id Disamina	Id Prodotto
39	2021-07-31 (31 Luglio 2021)	TD01 (fattura)	1121317686	N&TC SRL	1788.18	1555	1
39						1556	1
39						1557	2
39						1558	2
39						1559	3
39						1560	3
39						1561	4
39						1562	4
39						1563	5
39						1564	5
39						1565	6
39						1566	7
39						1567	7
39						1568	8
39						1569	8
39						1570	9
39						1571	9
39						1572	10
39						1573	10
39						1574	11
39						1575	11
39						1576	12
39						1577	12
39						1578	13
39						1579	13
39						1580	14
39						1581	14
39						1582	15
39						1583	15

Modifica dati fattura

### Modifica dati acquisiti

Id fattura: 39 - Id prodotto registrato: 1

**Dati generali del documento:**

Data documento: 2021-07-31 (31 Luglio 2021)

Tipologia documento: TD01 (fattura)

Numero documento: 1121317686

Denominazione: N&TC SRL

Importo totale documento: 1788.18

**Dati prodotti in fattura:**

ID	Prodotto/Sconto	Quantità	Val totale	IVA (%)	
1555	LATTE FRESCO ALTA QL	13.00000000	24.83	4.00	
1556	Sconto 58.00 % Linea 1	1.00	-14.40	4.00	
ID					
ID					
ID					

Così, nell'esempio dello screenshot appena mostrato, avendo selezionato la fattura 39 viene applicato un filtro su tutti i dati relativi a questa fattura... ma non solo.... Avendo selezionato la prima riga, essa risulta relativa all'id prodotto n°1, che è riportato due volte, con id disamina 1555 e 1556.... Essendo i valori selezionati, in fase di modifica saranno gli unici a poter essere modificati.

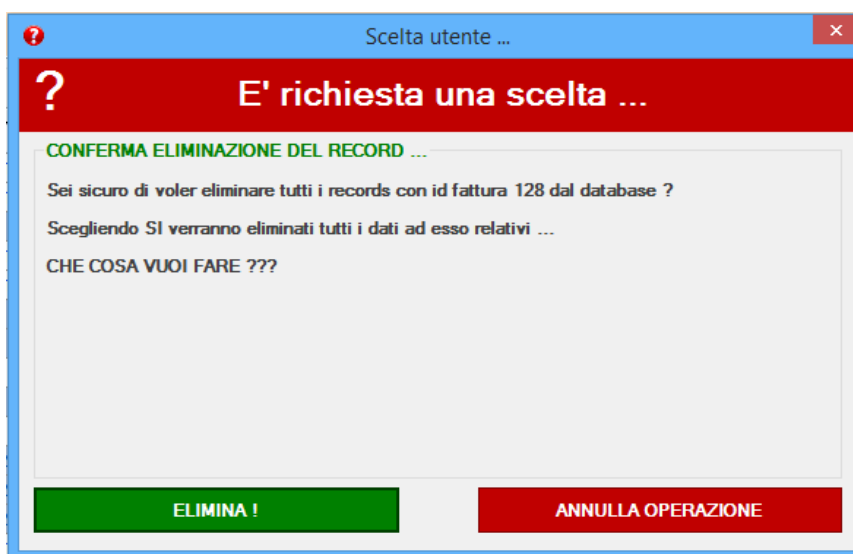
# PdfToDb, Estrazione di dati da PDF e salvataggio su database

Sviluppato da Stefano Ravagni

V 1.3.1.0 – Marzo 2023

## Elimina documento:

La voce elimina documento cancella dal database i soli dati relativi ad un preciso ID FATTURA selezionato.



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 128 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.

## Ricarica dati sulla griglia:

Serve semplicemente a forzare il caricamento dei dati estratti sulla griglia.

## Inserisci campo speciale:

Questa voce consente l'inserimento di un campo speciale valido per la sola disamina degli articoli... Il lancio di questa funzione comporta l'apertura di una apposita form per l'inserimento del campo speciale, che accetta valori alfanumerici fino a 254 caratteri.

Il campo speciale prende la denominazione impostata nelle impostazioni (si veda apposita sezione). Inserendo un valore e premendo il bottone per il salvataggio, il campo speciale sarà consolidato sul database e provocherà il ricaricamento della griglia dati della disamina, dove il nuovo valore sarà subito visibile.

# PdfToDb, Estrazione di dati da PDF e salvataggio su database

Sviluppato da Stefano Ravagni

V 1.3.1.0 – Marzo 2023

Di seguito lo screenshot della form.

Modifica dati del campo speciale

## Modifica dati del campo speciale

**Id fattura: 40 - Id prodotto registrato: 1**

**Dati generali del documento:**

Data documento:	2021-07-31 (31 Luglio 2021)
Tipologia documento:	TD01 (fattura)
Numero documento:	1121317686
Denominazione:	COOPERLAT SOC. COOP. AGRICOLA
Importo totale documento:	1788.18

**Barcode**

**45788**

 **Salva modifiche**

## Svuota tabella:

Previa conferma da parte dell'utente, svuota completamente la tabella relativa alla griglia selezionata.

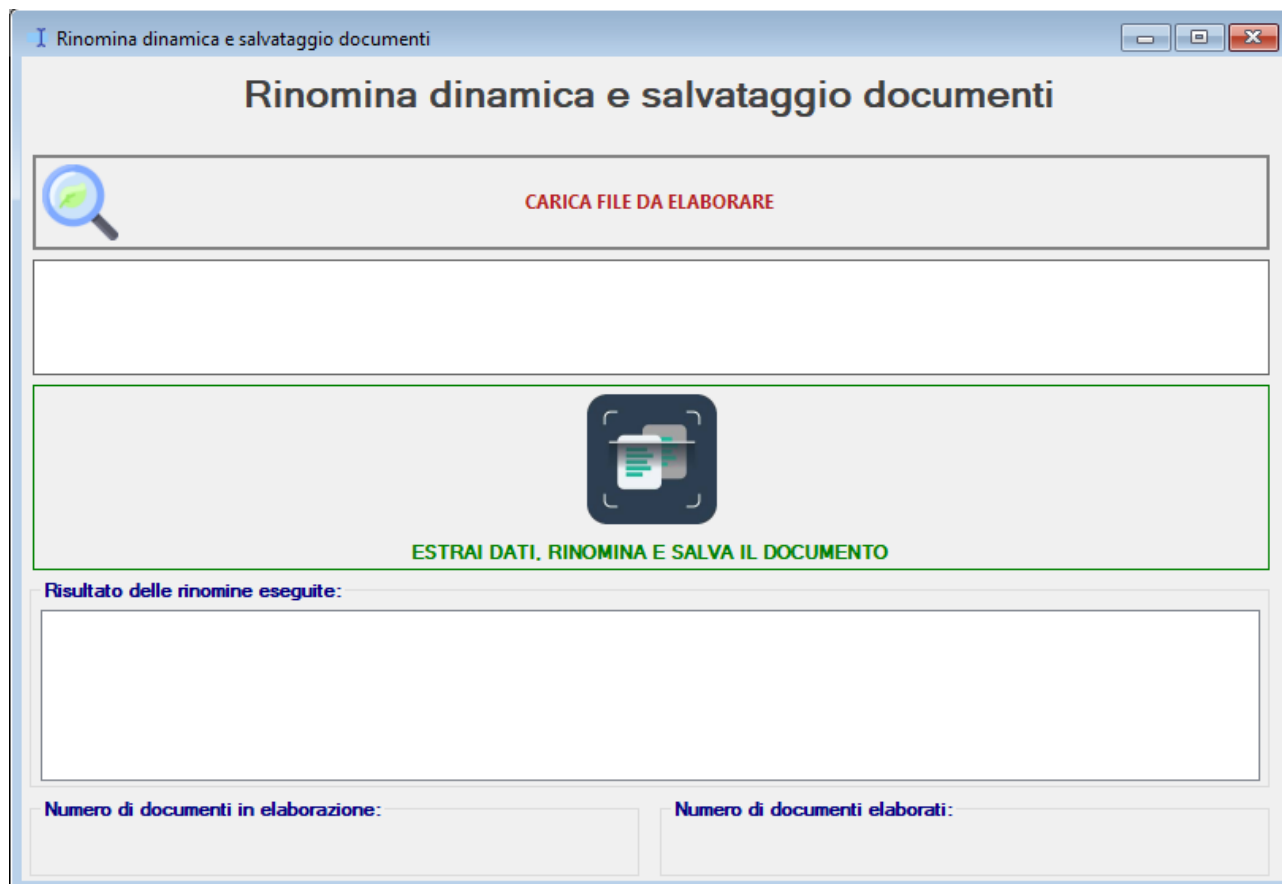
Al termine del processo di svuotamento viene lanciata la funzionalità di compressione del database Ms Access al fine di recuperare spazio e re inizializzare gli indici dei record che al successivo utilizzo ripartiranno dal numero 1.

# PdfToDb, Estrazione di dati da PDF e salvataggio su database

Sviluppato da Stefano Ravagni

V 1.3.1.0 – Marzo 2023

## Rinomina documenti



La schermata di rinomina è piuttosto semplice e presenta pochi controlli:

- Bottone **CARICA FILE DA ELABORARE**: consente di caricare i file pdf da rinominare, permettendo di sceglierli da qualsiasi sorgente.
- Bottone **ESTRAI DATI, RINOMINA E SALVA IL DOCUMENTO**: premendo questo bottone la rinomina dei file caricati sarà avviata... il risultato delle elaborazioni viene presentato nella list box **RISULTATO DELLE RINOMINE ESEGUITE**, dove viene riportato il nome del file originale e di quello rinominato. I file rinominati vengono salvati in automatico nel percorso indicato nella schermata delle impostazioni.
- Nei group box in fondo alla schermata è possibile vedere il calcolo dei file caricati e di quelli effettivamente elaborati, che ovviamente in condizioni normali deve corrispondere.

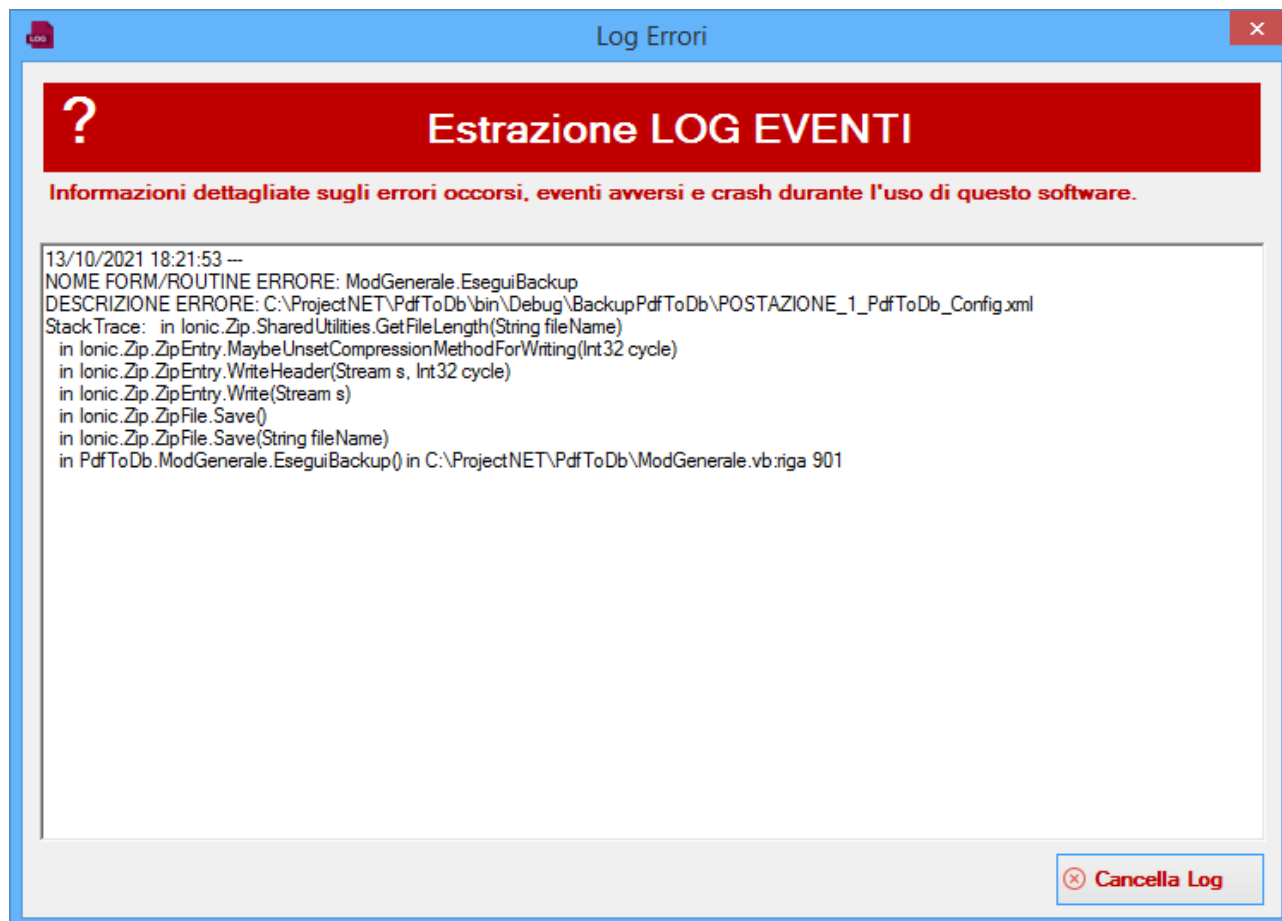


# PdfToDb, Estrazione di dati da PDF e salvataggio su database

Sviluppato da Stefano Ravagni

V 1.3.1.0 – Marzo 2023

## 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.

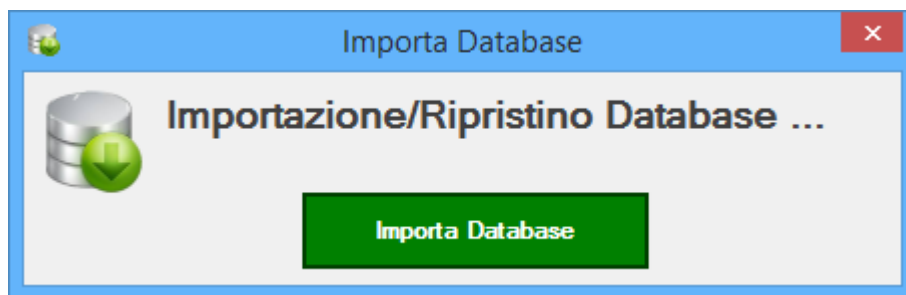
L'utente può cancellare il file di log con la semplice pressione del bottone CANCELLA LOG.

## Importazione database

# PdfToDb, Estrazione di dati da PDF e salvataggio su database

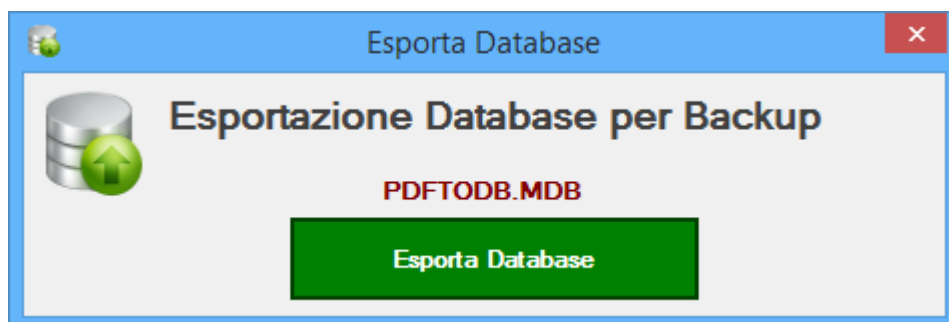
Sviluppato da Stefano Ravagni

V 1.3.1.0 – Marzo 2023



L'importazione del database è una funzione utile per ripristinare una base dati da 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 avvisa l'utente della necessità di questo pacchetto in fase di installazione.

Per chi avesse necessità di installarle sappia che le librerie di RUNTIME necessarie possono essere scaricate al seguente URL:

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

# PdfToDb, Estrazione di dati da PDF e salvataggio su database

Sviluppato da Stefano Ravagni

V 1.3.1.0 – Marzo 2023

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, per la sua gentilezza e professionalità dimostrata durante le fasi di sviluppo.