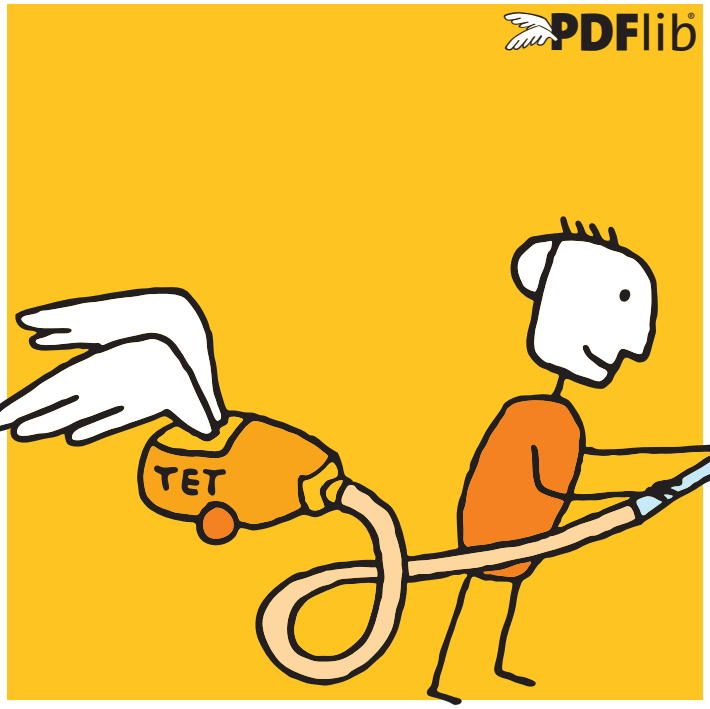


Datasheet

PDFlib TET 4

Text Extraction Toolkit

Che cos'è PDFlib TET?

PDFlib TET (Text Extraction Toolkit) estrae in modo affidabile testi, immagini e metadati dai documenti PDF. TET estrae il testo come stringa Unicode insieme ad informazioni dettagliate sui glifi, sui font e sulla posizione dell'immagine. Le immagini raster vengono estratte in formati raster comuni. Opzionalmente TET converte i documenti PDF in un formato XML chiamato TETML contenente testo, metadati e informazioni sulle risorse.

TET implementa degli algoritmi avanzati di analisi dei contenuti per identificare i limiti delle parole, raggruppare il testo nelle colonne e rimuovere il testo ridondante. Con l'interfaccia pCOS integrata è possibile recuperare qualsiasi oggetto dai PDF, come metadati, elementi interattivi, ecc.

Con PDFlib TET è possibile:

- Realizzare un indice PDF per i motori di ricerca
- Riutilizzare i testi e le immagini dei PDF
- Convertire i contenuti dei PDF in altri formati
- Elaborare i PDF in base al loro contenuto, ad esempio dividerli in base ai titoli (oltre a TET richiede PDFlib+PDI)

Funzionalità di PDFlib TET

PDF supportati

TET supporta tutti i principali tipi di PDF:

- Tutte le versioni PDF fino ad Acrobat 9, compreso ISO 32000-1
- PDF protetti che non richiedono una password per la consultazione
- I PDF corrotti vengono riparati

Unicode

Sebbene il testo nei PDF di norma non sia in formato Unicode, PDFlib TET permette la conversione in tale formato:

- TET converte il testo in Unicode. In C e in altri linguaggi che non supportano Unicode il testo sarà estratto in UTF-8 o UTF-16, mentre come stringhe native nei linguaggi con supporto Unicode.
- Le legature e gli altri simboli verranno tradotti nei corrispondenti caratteri Unicode.
- I simboli senza una diretta corrispondenza in Unicode vengono riconosciuti e convertiti in un carattere impostabile. In questo modo è possibile prevenire gli errori.
- TET implementa diverse soluzioni per ovviare ai problemi con specifici pacchetti, come documenti InDesign e TeX o PDF generati su sistemi mainframe.

Analisi dei contenuti ed identificazione delle parole

TET include strumenti avanzati per l'analisi del testo:

- Algoritmo brevettato per identificare i limiti delle parole e recuperarle
- Riunisce le parole suddivise in più righe (dehyphenation)
- Rimuove i duplicati, come le ombre ed il grassetto artificiale
- Ricombina i paragrafi
- Riordina correttamente il testo suddiviso nelle pagine

Layout della pagina e rilevazione delle tabelle

Il testo della pagina viene analizzato per rilevare le colonne di testo. Le tabelle vengono rilevate, comprese le celle che si estendono su più colonne. In questo modo viene migliorata l'estrazione del testo. Le righe della tabella e i contenuti di ogni cella possono essere identificati.

Geometria

TET permette la misurazione precisa del testo, così come la posizione nella pagina, le dimensioni dei simboli e la direzione di scrittura. Aree particolari possono essere incluse o escluse dall'estrazione, come i margini, gli header e i footer.

Estrazione immagini

Dai PDF è possibile estrarre file TIFF, JPEG e JPEG 2000. Per ogni immagine vengono indicate informazioni dettagliate (posizione, dimensioni e angoli). Le immagini frammentate vengono unite in immagini di più grandi dimensioni per semplificare il loro riutilizzo. Le immagini vengono riprodotte in modo fedele in quanto non viene eseguita la ricomputazione del file e la conversione degli spazi colore. In questo modo si garantisce il più alto livello di fedeltà possibile.

Analisi

TET include l'interfaccia pCOS per gestire tutti i dettagli del documento, come informazioni sul documento, metadati XMP, font, dimensioni e molto altro (vedi datasheet dedicato a pCOS).

Opzioni di configurazione per i PDF più impegnativi

TET prevede delle soluzioni specifiche per quei documenti PDF dai quali non è possibile estrarre il testo correttamente con altri prodotti. Inoltre, prevede diverse funzionalità di configurazione per migliorare l'elaborazione dei documenti complessi.

- La conversione in Unicode può essere personalizzata mediante una tabella contenente i codici carattere o il nome dei simboli in Unicode.
- PDFlib FontReporter è uno strumento ausiliario per analizzare i font, le codifiche e i glifi nei file PDF. Si installa come plugin per Adobe Acrobat. Per Mac e Windows il plugin è gratuito.
- I font embedded vengono analizzati per migliorare la conversione in Unicode. Se il font non è compreso, per migliorare la qualità dell'estrazione del testo vengono utilizzati i file dei font esterni o i font di sistema.

Unicode Postprocessing

TET supporta una serie di fasi di postprocessing in formato Unicode che possono essere richiamate per migliorare la qualità del testo estratto:

- Mantenimento del folding, rimozione o sostituzione dei caratteri, ad esempio rimozione di punteggiatura o caratteri da script non rilevanti.
- Decomposizioni per sostituire un carattere con una sequenza equivalente di uno o più caratteri, ad es. sostituzione di caratteri giapponesi stretti, larghi o verticali oppure delle varianti di superscript Latin (ad es. ª) con le rispettive controparti standard.
- Il testo può essere convertito in tutte le quattro forme normalizzate Unicode, ad es. emit NFC per soddisfare i requisiti per testo Web o per le basi di dati.

Aree del documento

I documenti PDF possono contenere del testo non solo nelle aree di contenuto. Sebbene la maggior parte delle applicazioni intervenga soltanto sui contenuti delle pagine, in molti casi possono risultare utili anche le altre aree del documento. TET è in grado di estrarre il testo dai seguenti domini:

- Contenuti della pagina
- Campi informativi predefiniti e personalizzati
- Metadati XMP a livello di immagine e documento
- Segnalibri
- I file allegati e portfolio PDF possono essere elaborati ricorsivamente
- Campi di moduli
- Commenti (annotazioni)
- Rilevamento delle proprietà generali dei PDF, come numero di pagine, conformità agli standard come PDF/A o PDF/X, ecc.

Metadati XMP

TET supporta i metadati XMP in diverso modo:

- Servendosi dell'interfaccia integrata pCOS è possibile estrarre in modo mirato i metadati XMP per il documento, le singole pagine, le immagini o altre parti del documento.
- L'output TETML contiene il documento XMP e, se presenti nel filePDF, i metadati delle immagini.
- Le immagini estratte nei formati TIFF o JPEG contengono i metadati, se questi sono presenti nel file PDF.

TETML rappresenta i contenuti del PDF nel formato XML

TET è in grado di rappresentare il contenuto dei PDF in un formato XML chiamato TETML. Questo formato contiene una grande varietà di informazioni in una forma che può essere facilmente elaborata con i più comuni tool per XML. TETML contiene il testo e, come opzione, le informazioni su font e posizione, i dettagli sulle risorse (font, immagini, spazi colore) e i metadati.

TETML si basa su uno schema XML che garantisce la creazione di file XML uniformi e affidabili. TETML può essere elaborato con gli stylesheet XSLT, ad esempio per applicare dei filtri o per convertire il file TETML in altri formati. La distribuzione TET comprende dei modelli di stylesheet XSLT per elaborare i file TETML.

Nel seguente esempio è possibile vedere una porzione di TETML con dettagli sui glifi:

```
<Word>
 <Text>PDFlib</Text>
 <Box llx="111.48" lly="636.33" urx="161.14" ury="654.33">
 <Glyph font="F1" size="18" x="111.48" y="636.33" width="9.65">P</Glyph>
 <Glyph font="F1" size="18" x="121.12" y="636.33" width="11.88">D</Glyph>
 <Glyph font="F1" size="18" x="133.00" y="636.33" width="8.33">F</Glyph>
 <Glyph font="F1" size="18" x="141.33" y="636.33" width="4.88">l</Glyph>
 <Glyph font="F1" size="18" x="146.21" y="636.33" width="4.88">i</Glyph>
 <Glyph font="F1" size="18" x="151.08" y="636.33" width="10.06">b</Glyph>
 </Box>
 </Word>
```

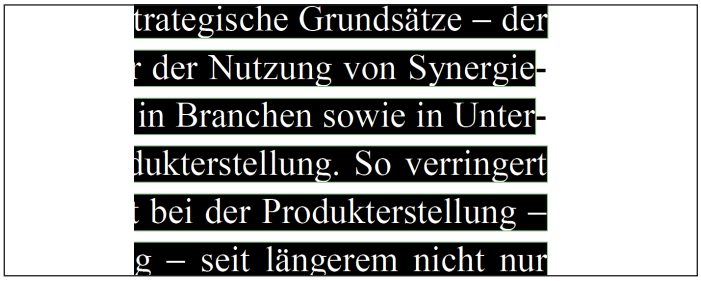
Interfacce TET

Le interfacce TET includono il codice necessario per interfacciare TET con altri programmi. Le seguenti interfacce TET rendono disponibile la funzionalità di estrazione del testo in numerosi ambienti software:

- Interfaccia TET per Lucene Search Engine
- Interfaccia TET per Solr Search Server
- Interfaccia TET per Oracle Text
- Interfaccia TET per MediaWiki
- TET PDF IFilter per i prodotti Microsoft è disponibile come prodotto dedicato. Estrae testo e metadati dai documenti PDF rendendoli disponibili ai programmi di ricerca su Windows (per maggiori dettagli fare riferimento al datasheet dedicato).

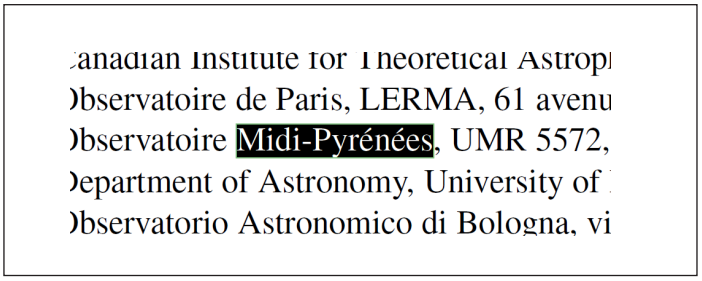
TET Cookbook

TET Cookbook è una raccolta di programmi dimostrativi che illustrano le varie possibilità di impiego di TET. I vari esempi mostrano come combinare i prodotti TET e PDFlib+PDI per elaborare e migliorare i documenti PDF, aggiungendo ad esempio dei bookmark o dei link basati sul testo presente nella pagina.

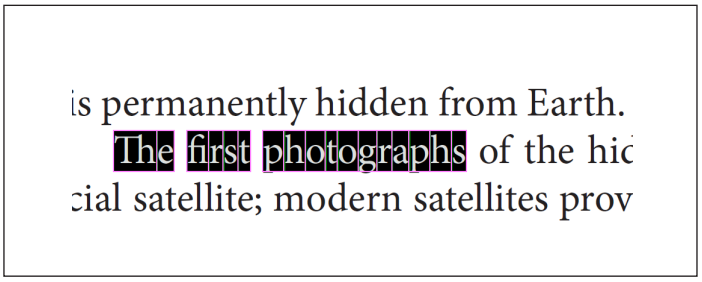


TET rimuove correttamente i trattini ma mantiene le lineette.

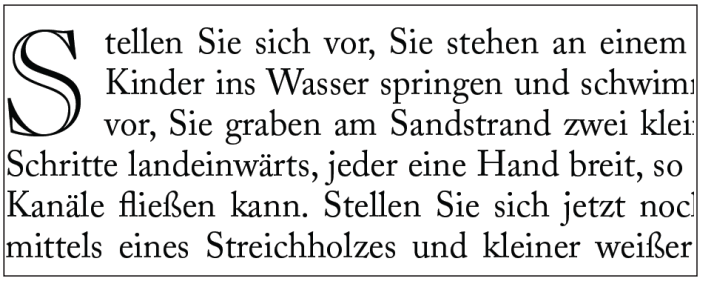
Altri prodotti estraggono »Inttrroduccttiion«.
TET estrae correttamente »Introduction«.



Altri prodotti estraggono »Midi-Pyr´en´ees«.
TET estrae correttamente »Midi-Pyrénées«.



Altri prodotti estraggono » e me fotografie«.
TET estrae correttamente »Le prime fotografie«.



Altri prodotti estraggono due parole: il capolettera »O« e »ggi«.
TET estrae correttamente la singola parola »Oggi«.

Le peculiarità dell'estrazione di testo da PDF

Parole divise da trattino

TET rileva le parole divise da trattino su linee multiple, rimuove il trattino e unisce le singole parti per formare una parola intera. Questa operazione è importante per permettere di ricercare le parole anche quando nei documenti queste sono separate da un trattino. Le lineette (differenti dai trattini per andare a capo) vengono trattate diversamente in quanto non devono essere rimosse.

Rilevazione di ombre o di grassetto artificiale

Molto spesso i documenti digitali contengono del testo ombreggiato. Questo effetto viene realizzato posizionando sulla pagina diversi strati di testo con un offset basso tra le singole parti. Allo stesso modo, il grassetto viene spesso simulato imprimendo più strati di testo l'uno sopra l'altro. Ne consegue che il documento contiene più istanze dei caratteri delle parole ombreggiate o in grassetto. L'algoritmo brevettato TET per il riconoscimento delle ombre identifica e rimuove le parti di testo ridondanti, evitando così l'estrazione eccessiva di testo. Mentre gli altri programmi estraggono più volte le parti di testo ombreggiate o in grassetto, TET rimuove correttamente le copie ridondanti. Anche se le parole doppie continuano ad essere identificabili durante una ricerca, i caratteri doppi dell'esempio impedirebbero la rilevazione della parola.

Caratteri accentati

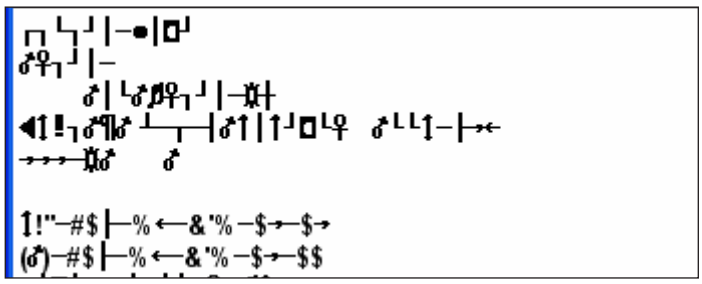
In molte lingue gli accenti o gli altri caratteri diacritici vengono combinati con altri caratteri. Alcuni programmi di tipografia, ad esempio TeX, producono due caratteri (caratteri di base e accento) separati per creare un carattere combinato. Per creare, ad esempio, il carattere *ä* viene impostata sulla pagina per prima cosa la lettera *a* quindi, su di essa, la dieresi ¨. TET riconosce questa situazione e ricombina i due caratteri per formare il carattere accentato.

Legatura

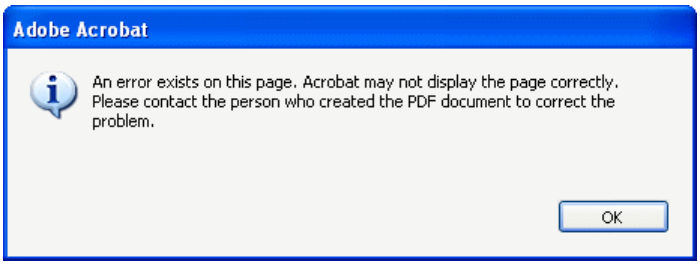
Una legatura è composta dall'unione di due o più lettere in un singolo glifo. Le legature più comuni vengono utilizzate nelle combinazioni *fi*, *fl* e *ffi*; alcune legature meno comuni vengono utilizzate per le combinazioni *Th*, *sp*, *ct*, *st*, ecc. Durante l'estrazione del testo dai documenti PDF è necessario analizzare le legature per separare i grafemi che le costituiscono ed elaborare correttamente il testo. TET rileva le legature e le traduce in due o più caratteri corrispondenti.

Capolettera

Con capolettera si intende un carattere di maggiori dimensioni con il quale inizia un paragrafo. Nella parte alta, il capolettera èallineato con la parte superiore della riga, il resto del carattere occupa invece diverse righe verso il basso. I capolettera sono utilizzati per mettere in rilievo l'inizio di un paragrafo. Se non vengono accuratamente trattati, la prima lettera verrà estratta in due parti: il singolo carattere iniziale e il resto della parola.



Mentre gli altri prodotti prducono un risultato inutilizzabile, TET fornisce testo.



I contenuti delle pagine non vengono visualizzati nemmeno da Acrobat, ma TET estrae correttamente il testo.

TET riordina la visualizzazione mista di testo da destra a sinistra e da sinistra a destra e crea un testo in uscita corretto.

Altri prodotti estraggono 133 minuscole parti.
TET estrae un'unica immagine.

Le peculiarità dell'estrazione di testo da PDF

Mappatura Unicode

La mappatura Unicode costituisce la premessa per l'estrazione di testo da PDF: ad ogni glifo presente nella pagina deve essere assegnato il corrispondente valore Unicode. Il formato PDF rende questo compito particolarmente arduo in quanto supporta una vasto numero di font e di codifiche senza fornire sempre le informazioni necessarie per assegnare il corretto valore Unicode. Nel peggiore dei casi, quando il documento non fornisce alcun tipo di informazione, il testo estratto dal documento sarà inutilizzabile.

TET si avvale di un algoritmo brevettato per la mappatura Unicode. Questo algoritmo "a cascata" sfrutta tutte le informazioni disponibili per determinare il valore Unicode. Anche con i documenti più problematici, quando i prodotti della concorrenza falliscono, TET è in grado di estrarre correttamente il testo Unicode.

Documenti PDF corrotti

I documenti PDF possono corrompersi ad esempio in caso di errori di trasmissione o altri tipi di problemi. La modalità di riparazione implementata in TET recupera molti tipi di PDF danneggiati. In alcuni casi, i file PDF sono così danneggiati da non poter essere aperti nemmeno da Acrobat. Anche in questi casi TET è molto spesso in grado di accedere ai contenuti della pagina.

Testi bidirezionali con arabo ed ebraico

PDF non codifica il testo logico, ma è un semplice contenitore dei glifi presenti nella pagina. In arabo e in ebraico il testo va da destra a sinistra. Poiché spesso contiene inserzioni da sinistra a destra, ad esempio nel caso di numeri o nomi in lingue occidentali, il testo deve essere interpretanto in entrambe le direzioni. Ecco perché si parla di »bidirezionale«. L'arabo pone delle ulteriori difficoltà, poiché i caratteri possono essere utilizzati in quattro forme diverse. Le varie forme che possono assumere i caratteri devono essere normalizzate nel rispettivo standard (isolato).

Le peculiarità dell'estrazione di immagini da PDF

Spazio colore e compressione

I dati delle immagini rasterizzate nei PDF possono essere codificate in una delle undici combinazioni di spazio colore e nove filtri di compressione. Tuttavia, i formati immagini più comuni come JPEG e TIFF ne supportano solo una sottoparte. L'estrattore immagini di TET bilancia in modo attento le caratteristiche dell'immagine PDF con le capacità del formato di uscita dell'immagine. Indipendentemente dalla struttura interna dell'immagine PDF, l'immagine in pixel verrà estratta in uno dei formati standard di immagine.

Deframmentazione delle immagini

Le immagini contenute in moli PDF sono suddivise in più parti dalle applicazioni che hanno prodotto il PDF. Quello che sembra un'unica immagine sulla pagina in realtà può essere formato da centinaia di migliaia di piccole parti. Tra gli altri, Microsoft Office e TeX producono immagini di questo tipo. TET rileva le immagini frammentate e unisce le varie parti in un'immagine unica facilmente riutilizzabile. È solo grazie alla funzione di deframmentazione che risulta possibile continuare ad usare l'immagine.

Gli usi di TET

TET è disponibile come libreria per diversi ambienti di sviluppo e come command line per le operazioni di batch. Entrambe le versioni offrono le stesse funzionalità, ma sono indicate per scenari differenti. La libreria TET e il command line TET sono in grado di creare TETML, il formato XML di TET.

TET offre le seguenti opzioni di utilizzo:

- La libreria di programmazione TET (componente) è adatta per essere integrata nelle applicazioni desktop o server. Diversi esempi su come utilizzare la libreria sono disponibili nel pacchetto TET.
- La command line TET è ideale per l'elaborazione batch dei documenti PDF. Non richiede nessun tipo di programmazione ma offre delle opzioni che permettono di integrare l'applicazione anche nei workflow più complessi.
- Il formato TETML è adatto per i workflow basati su XML e per i programmatori che operano con questo formato e che conoscono i numerosi strumenti e i linguaggi di elaborazione dei file XML, come XSLT.
- Le interacce TET sono adatte ad integrare TET in vari pacchetti software, ad esempio base dati e motori di ricerca.

La famiglia di prodotti TET

TET comprende i seguenti prodotti:

- Il pacchetto TET principale così come descritto nel presente datasheet.
- TET PDF IFilter è disponibile come prodotto separato. È adatto per essere utilizzato con i prodotti di ricerca della Microsoft, ad es. Windows Search, SharePoint e SQL Server (per maggiori dettagli fare riferimento al datasheet dedicato).
- Il plugin TET per Adobe Acrobat è un'utility gratuita per estrarre testo e immagini da PDF. Può essere usato per valutare TET in modo interattivo.

Ambienti di sviluppo supportati

PDFlib TET funziona su quasi tutte le piattaforme informatiche. Offriamo versioni 32-bit e 64-bit per tutte le più comuni versioni di Windows, Mac OS, Linux e Unix, nonché per i sistemi IBM i5/iSeries e zSeries.

Per garantire le migliori performance possibili e ridurre l'overhead, il TET core è scritto in codice C ottimizzato. Grazie ad una semplice API (Application Programming Interface) è possibile accedere alle funzionalità di TET da numerosi ambienti di sviluppo:

- COM per VB, ASP, Borland Delphi, ecc.
- C e C++
- Java, servlets e Java Application Server
- .NET per C#, VB.NET, ASP.NET, ecc.
- Perl
- PHP
- Python
- REALbasic
- RPG (IBM i5/iSeries)

Benefici di PDFlib

Affidabile

In tutto il mondo sono migliaia i programmatori che utilizzano il nostro software. PDFlib soddisfa tutti i requisiti in termini di qualità e performance per l'utilizzo sui server. Tutti i prodotti PDFlib sono adatti per l'utilizzo continuo su server e per le operazioni di batch senza supervisione.

Semplice e veloce

I prodotti PDFlib sono estremamente veloci e sono in grado di elaborare fino a centinaia di pagine al secondo. L'interfaccia di programmazione è intuitiva e di facile apprendimento.

I prodotti PDFlib sono utilizzati in tutto il mondo

I nostri prodotti supportano tutte le lingue internazionali e Unicode. Vengono utilizzati da utenti in ogni angolo del mondo.

Assistenza professionale

In caso di problemi potete contare sul nostro servizio di assistenza professionale. Per garantire che il workflow a livello aziendale non subisca mai interruzioni, offriamo la possibilità di sottoscrivere un contratto di servizio. In questo modo potrete accedere alle versioni più recenti e potrete contare su tempi di risposta celeri qualora si dovesse verificare un problema.

Licenza

Sono a disposizione diverse tipologie di licenza: server, integrativa, aziendale e codice sorgente. È altresì possibile sottoscrivere un contratto di servizio per usufruire di un'assistenza professionale con tempi brevi di risposta e aggiornamenti gratuiti.

L'azienda PDFlib GmbH

PDFlib GmbH è un'azienda specializzata nel settore delle tecnologie PDF. I prodotti PDFlib vengono utilizzati in tutto il mondo dal 1997. L'azienda segue attentamente gli sviluppi e i trend del mercato, come gli standard ISO per il formato PDF. PDFlib GmbH distribuisce i suoi prodotti in tutti i continenti e detiene una posizione leader nei mercati in Nordamerica, Europa e Giappone.

Contatti

Sul nostro sito Internet sono disponibili le versioni demo complete di documentazione ed esempi. Per maggiori informazioni, contattare:

PDFlib GmbH
Franziska-Bilek-Weg 9, 80339 Monaco, Germania
Te. +49 • 89 • 452 33 84-0, Fax +49 • 89 • 452 33 84-99
sales@pdflib.com
www.pdflib.com