

## Data entry system for large volumes of forms and unstructured documents

The screenshot displays the PAPERIN software interface. The main window is titled "Manager [deutsch]" and contains a menu bar with "Engine", "Eingabewarteschlange", "Ausgabewarteschlange", "Optionen", "Werkzeuge", and "Hilfe". Below the menu bar is a toolbar with various icons. The main area is divided into two panes: "Eingabewarteschlangen" (Input Queues) on the left and "Ausgabewarteschlangen" (Output Queues) on the right. The "Eingabewarteschlangen" pane shows a list of files named "arena-001.tif" through "arena-010.tif". The "Ausgabewarteschlangen" pane shows a list of files including "InterSport\_Average", "Intersport\_err", "Intersport\_ok", and "InterSport\_Export".

Overlaid on the main window is a smaller window titled "Form Recognition Client (v 4.47)". This window has a menu bar with "Start", "Stop", "Optionen", "Größer/Kleiner", and "Hilfe". It features a toolbar with icons for "OCR", "Kadmos", "ScanSoft", and "Beenden". The "OCR" section is active, showing "Dateiname: arena-002.tif" and "Formular: Arena". The "Dateien" section shows "Gut: 2" and "Mittel: 0". The main area of the "Form Recognition Client" displays a scanned document with various fields highlighted in red and green. The document is a receipt or invoice with the following text:

**Rechnung**  
 Kunde: 71402 Re-Nr.: 562407  
 Re-Def.: 22.02.01

005 Verb.: ISPO 507140 Rab.: 3.00%  
 Auftrag vom: 05.09.00 12 894  
 DER, BA: MOBI Seite 1

prozessieren Stck. Preis Wert  
 DEM DEM

ang vom: 22.02.01 Auftrag: 234364  
 LS: 144868

helmstraße 16  
 1638 Ludwigsburg  
 140

The "Form Recognition Client" also has a sidebar on the left with a list of fields and their values:

- Rechnungs Datum: 22.02.2001
- Rechnungs Nummer: 562407
- Blatt Nr: 1
- Lieferschein:
- Betrag Lieferung L: 19.61
- Betrag Porto L: 0.00
- Zwischen Summe L: 19.61
- Betrag MWST L: 3.14
- Betrag Summe L: 22.75

At the bottom of the "Form Recognition Client" window, there is a log of events:

```

11:10:14 C:\PaperIn4\Queues\Input
11:10:15 C:\PaperIn4\Queues\Input
11:10:15 C:\PaperIn4\Queues\Input
11:10:15 C:\PaperIn4\Queues\Input
11:10:32 arena-001.tif recognized , =
11:10:36 arena-002.tif recognized , =
  
```

## **PaperIn...**

- replaces manual input of documents for further processes
- interprets defined areas on forms
- interprets unstructured documents as well as their allocation to a workflow
- interprets faxes and allows to manage their content to defined processes.
- recognizes Latin and Cyrillic printed fonts with XIS (Kurzweil) ICR (Intelligent Character Recognition) processing without predefining the font family .
- descew of inclined scans or printed pages; precise positioning of defined fields according to graphical attributes on the document; removal of interfering lines, undesirable signs and noise.
- separates defined processes in different directories.
- based on a modular concept. Functions and parts of the program may be optimized to meet individual requirements with different working tasks. All modules use a similar screen layout and can infringe upon, extend to, and are complementary to one another.
- multilingual screen layout.

## **PaperIn...**

### **Server module:**

- **DocRec Manager:** Management of database with configurations, users, projects
- **FormrecWS:** Document recognition

### **Client Module:**

- **DocRec Designer:** Definition of forms and Documents
- **DocRec PostEdit:** Editor for corrections with image viewing
- **DocRec Admin:** Definition of users, projects and access rights

### **additional module:**

- **Analysis of documents:** Content related exploitation

### **Technical specifications:**

**Character recognition/fonts:** printed fonts (Omnifont<sup>1</sup>), Glyph<sup>3</sup>, check boxes, single character handwriting<sup>2</sup> and barcode

**Character size:** 8-72 dots at 200 dpi

**Area :** Unlimited number of fields, freely configurable

**Criteria:** Alphanumeric, numeric, monetary, date

**Validation:** Built in and user definable validation

**User interface:** Graphical menu driven user interface

**Form definition:** Interactive, according to scanned documents

**Preprocessing:** Removal of lines, boxes and noise, deskew of image

**Postprocessing:** Interactive, image related

**Batch processing:** Multiqueue management

**Image processing:** Saving of descewed images (optional)

**Image acceptance:** Various resolutions and formats, z.B. CCITT G3/G4

**Data output for text:** ASCII, ISO, Xdoc<sup>2</sup>, PDF<sup>3</sup>, XML

**For databases:** reference of image to outside area of database or embedded in file.

**Operating system:** NT 4.SP6<sup>4</sup>, Windows 2000<sup>4</sup>, Windows XP<sup>4</sup>

**PC requirements:** Intel Pentium III, min. 1GHz, min. 256 MB RAM, minimum SuperVGA; 19" monitor recommended

### **Producer:**

ARPA Data GmbH  
Albisstrasse 36

CH 8134 Adliswil

Tel: +41 1 709 09 79

<sup>1</sup>Registered trademark of Scansoft™

<sup>2</sup>Trademark of RE Recognition Technology GmbH

<sup>3</sup>trademark of Xerox Corporation

<sup>4</sup>Registered trademark of Microsoft™