



DOKuStar Extraction

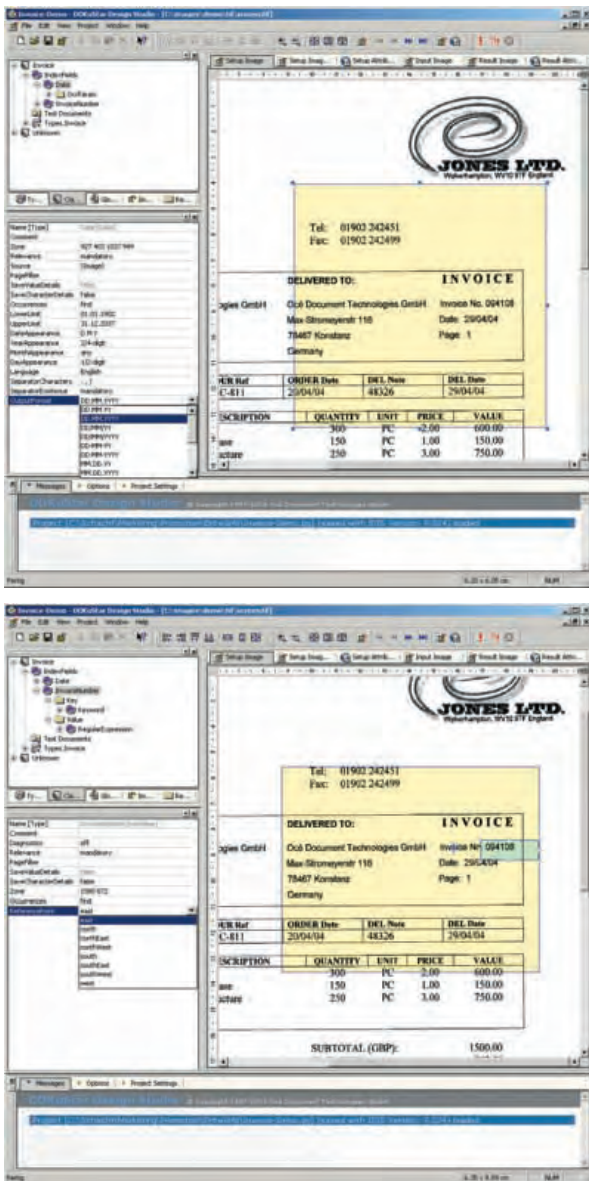


Automatic Classification, Indexing and Interpretation

DOKuStar Extraction is a software engine that extracts required data from scanned or coded documents. It is integrated into data capture systems, capture platforms, document management systems or dedicated applications such as SAP® to save users the tedious and error-prone job of keying in data and to automate document-based business processes. DOKuStar Extraction is supplied with input documents via a programming interface – the result is made available to the downstream application as an XML document.

DOKuStar Extraction offers OCR, ICR and document analysis in a single engine. Optical Character Recognition (OCR) converts pixels into characters whereas Intelligent Character Recognition (ICR) processes, for instance, handwritten information and complex forms. Document analysis is based on the results of character recognition and locates certain specific information such as the recipient of a letter and converts this information into Given name, Surname, Street, Number, Postcode and City.

Document analysis in DOKuStar uses free-form technology. This allows documents with very different layouts to be extracted without defining a separate template for each layout. This means that the application is more productive and also more robust when changes are made to the documents to be processed.



Field Types – Building Blocks of Document Analysis

Field types are central to data extraction with DOKuStar. A field type – such as Date – searches a prescribed zone of the document for date information. The zone can be as large as the whole document but it can also be as small as an input field in a form. Optical character recognition is undertaken for the defined zone – the data is then examined for structures with a date-like syntax.

The search is controlled by many different parameters. Whether only German, American or any date format is permitted, whether the month may be written in numerical or alphanumeric characters, or whether only a certain time period is permitted: All these factors can be parameterised to control the search process. The default parameter settings allow the detection of all date formats.

DOKuStar does more than search for a character sequence with a defined syntax in a character string. OCR results are never 100% perfect. Poor paper quality, poor image quality or dirt result in individual characters being questionable, unreadable, or wrong. DOKuStar takes these potential sources of error into account. The OCR can, for example, read an 'O' (capital letter) as a '0' (zero). DOKuStar knows this and can correct this error during date search. Here it is obviously of benefit if character recognition and document analysis software are developed by the same company.

Combining Field Types: The Whole Equals More than the Sum of the Parts

DOKuStar has ready-to-use field types for frequently encountered data types such as a date, amount or address. Application-specific data types such as order number can be modelled with so-called "regular expressions".

Normally there are several amounts or dates in a document, but only a specific one is needed, for example the order date. Combined field types such as the KeyValue field allow further delimitation of the search. The key can be a phrase or keyword, possibly from a long list; the value is the date field. DOKuStar searches for these specified phrases and selects a date located close to the phrase. If there are various possibilities for the phrase or date values in the document, DOKuStar selects the most plausible hypothesis. The FirstOf field type examines a series of hypotheses and selects the first applicable one. Combined with the KeyValue search, this allows powerful search procedures to be implemented quickly and simply.

User Definition of Field Types: DOKuStar Custom Operators

Processing complex search strategies with simple basic elements – that is the principle of DOKuStar Extraction. Once a search definition has been created, it can be used as an index field or classification criterion for different types of documents in any project.

Such definitions, which are called Custom Operators, can be stored and distributed like macros. The possibility of password-protecting Custom Operators allows the know-how intrinsic to a Custom Operator to be protected and marketed independently of DOKuStar. Complete application-specific libraries of field types can be established in this way.

Field Types	Description
Date	Finds and reads the date
Amount	Finds and reads the amount
Address	Finds and reads the address
RegularExp	Defines fields for which no templates are available
Text	Reads fields in a set position
Keyword	Finds and reads a keyword
Phrase	Finds and reads a phrase
Wordset	Finds value combinations on the document
KeyValue	Finds and recognises a value relative to defined field
FirstOf	Allows multiple attempts to locate a field
BarCode	Finds and determines all common one-dimensional barcodes
PixelCount	Determines the density in a specified field
CheckBoxGroup	Combines the results of several PixelCount fields
Table	Finds and reads a table
FuzzyDB	Performs a fuzzy match reconciliation with a database for imported data
InvoiceDate	Finds and reads the invoice date
InvoiceNumber	Finds and reads the invoice number
InvoiceOrderNumber	Finds and reads the order number
InvoiceVendor	Database-supported supplier identification
InvoiceTotals	Finds and reads the total and net sum, VAT rate for individual VAT categories and the invoice total field as well as additional costs

Image Processing Functions

The extensive image processing functions offered by DOKuStar can remove annoying elements such as lines, dirt or backgrounds from forms and typical business documents. The image processing functions can be used uniformly for all documents to be processed or specifically for certain document types or classes. Document-specific image processing is useful if the image processing functions used have a detrimental effect on documents of another type.

Document Classification

Document analysis with DOKuStar takes place in two steps: Document classification and data extraction. Classification assigns the document to a class, such as invoice, order, application form, business letter. Defined index fields are extracted for each document class.

The documents are classified as required by the target application. For example, if all incoming orders are processed in the same way, a single document class is sufficient. If, however, orders from affiliated companies are to be handled separately, these can be modelled as a separate document class.

It is also possible to classify documents more precisely, namely into document types. This allows specific templates to be defined for special documents – in addition to the free-form rules for data extraction – thus substantially enhancing recognition accuracy.

Typical classification criteria are keywords, phrases or logos. These are also available as field types and are used to specify classification rules. They can be combined with KeyValue and FirstOf fields.

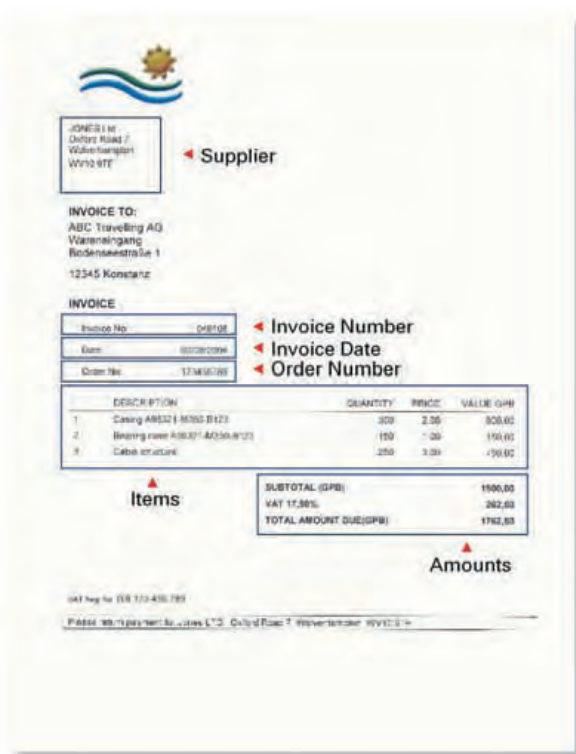
Extracting Data from Invoices: DOKuStar Invoice Option

Captaris Document Technologies offers a set of special field types for the extraction of invoice data. These can read the relevant fields – supplier, invoice date, invoice number, order number, amounts (additional expenses, net amount, tax, total amount, tax rate, currency) and the invoice items – from a random mix of invoices. The search strategy for this data has been optimised based on a huge number of national and international invoices.

DOKuStar Design Studio

The DOKuStar Design Studio is the development and test environment for DOKuStar.

The document classes and index fields for an application are defined in the Design Studio. The setup procedure uses document types, image processing functions, field types and custom operators. Frequently used definitions can be stored as global definitions. The result is a project file that controls the DOKuStar engine when processing documents in server mode.



Example images and test batches can be defined for each document type. Batch runs detect missing or erroneous search definitions quickly and reliably. Short test and change cycles ensure that the application is optimised in a short time.

Interfaces

DOKuStar Extraction is integrated into the target application via a programming interface. The extraction results are available to the application as an XML file.

© 2008 All rights reserved. No part of this publication may be reproduced, transmitted, transcribed, stored in a retrieval system, or translated into any language in any form by any means without the written permission of Captaris. All brand names and trademarks are the property of their respective owners.
Order Nr. L93061-N140-X-2-7618, 415 178, 02/08, printed in Germany.

Headquarters (Germany)
Captaris Document Technologies GmbH
Max-Stromeyer-Str. 116
78467 Konstanz

Phone (+49) 7531 87-4500
Fax (+49) 7531 87-4567
www.Captaris-dt.com

US Office
Captaris, Inc.
Captaris Document Technologies
4340 East West Highway, Suite 201
Bethesda, MD 20814
Phone (301) 652 9732
Fax (301) 652 7088
www.Captaris-dt.com