

INTELLIGENT DOCUMENT RECOGNITION – HOW DO YOU RUN DOCUMENT CAPTURE WITHOUT IT?

Intelligent Document Recognition, or IDR, is set to revolutionise the way in which we capture and process documents.

Today's IDR applications are focusing on invoice processing and similar applications involving semi-structured documents. Tomorrow, the techniques and technologies will be deployed to address key applications such as mailroom automation, and automatic indexing and classification of scanned documents.

So, what IS IDR? How does it work? What are the critical success factors? And why should you be thinking about it today?

What IS IDR, and how does it work?

IDR systems analyse the topology and the content of documents to make intelligent assessments of document type and key data required.

Example: if an IDR application finds the word "invoice" on a document, it thinks it might be an invoice. If it has some combination of "invoice number:", "invoice date", "total", "sub-total", "VAT", "item", quantity" "unit price" and other clues, appearing in more or less the right place on the document, then the system can be fairly sure it's an invoice.

A rules engine is used to determine which data needs to be extracted, for the purposes of indexing, business process input, repository management, or input into a knowledge-based application such as CRM.

IDR systems use a range of techniques for document analysis. These include:

- ▼ Automatic classification engines, using natural language processing and/or probabilistic techniques.
- ▼ Proximity – This number is close to the word "total", so it is probably the total.
- ▼ Document Layout – these words or characters appear in a vertical line down the page, under what looks like a heading of "item", so this might signify a set of line items on an invoice.
- ▼ Validation, dictionary matching and look-up techniques – this looks like the right format for one of our account numbers – what name, address, etc. do we have against that account number? Do they match with what we think is the name and address in the document?

Underlying the whole thing is a workflow-type system, which understands the business rules and the processes and handles exception processing, document and data routing, and audit trails etc.

Why is all this so important?

IDR systems deliver a number of key benefits that are simply not available elsewhere.

For one thing, what about the benefits of automatic indexing? How many high-volume document imaging systems have failed because of the indexing overhead? How many document repositories are lying unused and unusable because a change of circumstance has rendered the existing index schema inappropriate?

Mailroom automation systems should be able to differentiate between a letter of complaint, an invoice, an application form, a CV, and treat each one appropriately.

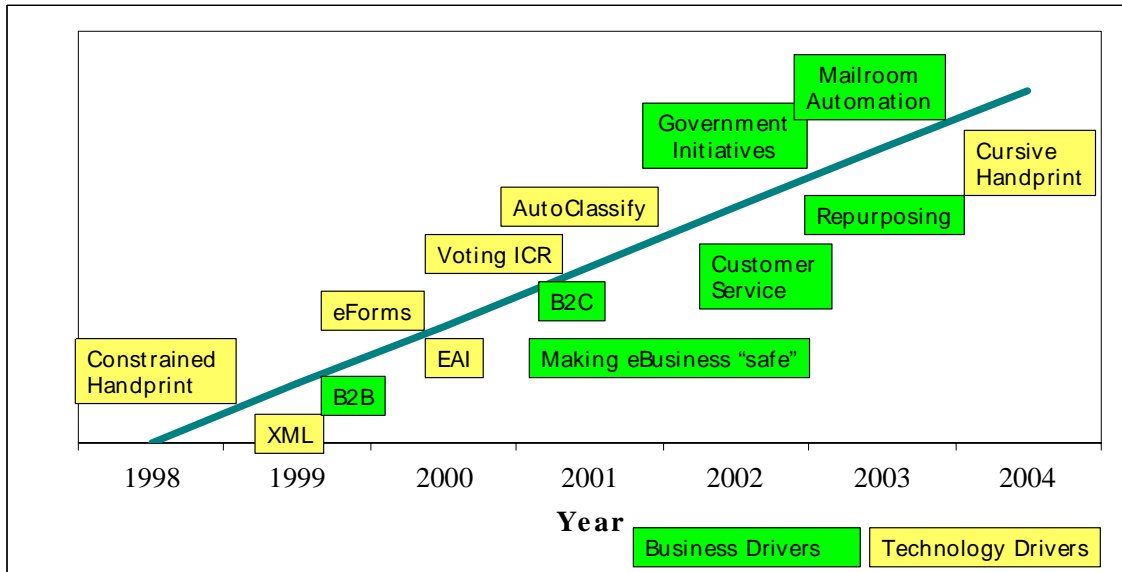
IDR applies to forms processing applications in two main ways. Where it's not possible to dictate document layout (as it is with forms you send out for completion), there is simply no other way. IDR also removes the overhead of having to re-specify the form every time a small change is made. As such, IDR empowers change and business agility.

Some key issues to consider

One issue is capture and recognition quality. Even the best IDR technology will fail unless it is presented with good quality information from the captured document. Advanced recognition capabilities, including multi-pass recognition through multiple engines, combined with intelligent analysis of the results, document clean-up facilities including deskewing, line and background removal, colour processing and the development of cursive handwriting recognition technologies – all play a part in ensuring good performance of such systems.

The biggest single mistake that is constantly made, however, is in how these systems are bought and sold. Vendors oversell; buyers entertain unrealistic expectations and are disappointed when the systems don't deliver. Business cases have to be built around real-world results and the business benefits they accrue. In cheque processing, for example, it is proven that 60-70% success rates deliver significant benefits. If our ICR does not achieve 99.99% accuracy, then we feel disappointed.

Figure 1. Drivers for Intelligent Document Recognition



The bottom line

Document capture systems that do not use IDR technologies for forms processing and/or automatic indexing and classification of scanned documents will be, in relative terms, obsolete by end 2005. The choice is yours...

Author: John Richardson. Projects Director of the independent market analysis, research and consulting firm Strategy Partners, and lead author of "Document Capture for eBusiness 2001-3".

Copyright ©2002 Strategy Partners International Ltd. All rights reserved

For all information on Strategy Partners services and publications contact:

advice@strategy-partners.com

Or visit <http://www.strategy-partners.com>