

Intelligent Document Identification

Reduce or Eliminate the Expense of Manual pre-Sorting
and Insertion of Separator or Barcode Pages

An Introduction to *Wordfire* Classify



January 2008

Datacap Inc
660 White Plains Road
Tarrytown, NY 10591

Table of Contents

Executive Summary..... 2

The Document Identification Challenge 3

 Document Identification Methods.....3

Wordfire Classify to the Rescue 5

 Provides Human-like Cognitive Recognition.....5

 Exception Processing6

 Unstructured Documents6

 Distinguishing Features6

How it Works 7

 The Learning Process.....8

 Production Use.....9

Summary 10

Benefits..... 10

About Datacap 10

Executive Summary

The most critical step in the process of automating document processing is to properly *identify* each and every document. The scanner hardware knows nothing about the documents being scanned. It just creates image files. You need to identify each document so that it can be correctly processed. There are two fundamental methods of identifying a document; either you have an operator identify the document prior to scanning, or use document capture software to identify the document from the image file.

Having the operator place a document separator page between individual documents, or separate documents into piles of different document types, is a manual and therefore potentially error-prone process. It slows down your ability to rapidly process the documents your business depends on, putting you at a competitive disadvantage.

This paper examines the challenges of document identification and presents five traditional methods that software uses to identify documents. It then introduces *Wordfire Classify*, a unique solution that identifies documents in a new way, using text analytics for true human-like content analysis. *Wordfire* is the closest document identification solution to human cognitive recognition. You will learn how this unique new product works and how you will benefit by using it.

Once you know what the document is, you can begin to process it in earnest. You can extract key information for indexing and send the information and the document image to the appropriate application or workflow system. Datacap Taskmaster, in combination with *Wordfire*, provides the most complete set of document identification capabilities available in a single product on the market today.

The Document Identification Challenge

The route to the paperless office begins with paper. Even today, many business interactions require paper: process a mortgage application, bind an agreement with a contract, complete a survey, submit an application, process a loan, make a medical claim, or fill out an ID card. In all these cases, we complete something on paper and then send it in.

Some paper must be manually handled and delivered to the recipients in your organization. But the paper documents that are critical to your business, those from your customers or prospective customers, documents such as mortgage application packages or legal depositions, must be quickly and accurately processed. These documents fuel the business activities that drive your success. And time really is money. If you have a person read each and every business-critical document coming into your organization to identify what it is, in order to know who to send it to, you will be at a competitive disadvantage. These are the paper documents you need to turn into electronic documents and process at machine speeds.

All enterprises seek to improve efficiency, accuracy, and the speed of document processing. Mortgage loan originators receive a large volume of customer correspondence and documents such as notes, insurance certificates, disclosures, waivers, truth-in-lending statements and payment letters. Insurance companies must process claim forms, proposals, correspondence, signature forms and contracts. Legal firms, healthcare organizations, financial institutions and governments all process large volumes of paper. Any organization receiving mixed document types must find an efficient and accurate way to identify each individual document.

Is it an order form, a complaint letter, a policy, a statement, affidavit or other document type? Once you know what it is you know what to do with it. Today, some organizations employ people to manually read and sort these inbound documents before they are scanned and routed for proper action. Other organizations have automated some portion of document identification, manually processing the remaining documents. Days can pass before documents are manually reviewed, sorted, and properly routed. Further delays are caused by incorrect manual document identification. The holy grail of document identification is to find a software capture solution that automates identification for 100% of your documents.

Document Identification Methods

The software used for document - and forms-capture provides multiple methods of identification. One goal of this software is to use the minimum computer resource necessary to properly identify a document. This helps make document identification fast and efficient. Some documents are easy to identify because they have a fixed structure. The capture software can focus on a particular location, or locations (often called zones) on the document and look for identifying logos, images or text. In these cases, the whole document may not have to be read, which means less processing compared to reading and searching an entire document page. Most capture and forms processing software support the following methods of document identification:

- **Optical Character Recognition (OCR) or Intelligent Character Recognition (ICR):** Identifies machine print or handprint documents based on text keywords or phrases. The best capture software lets you choose between different OCR or ICR engines, as appropriate.
- **Optical mark recognition (OMR):** Identifies documents based on checkboxes, bubbles, or other shapes that have been filled in by hand.
- **Bar code recognition:** Identifies documents based on finding regular bar codes, or high-density two dimensional (2D) bar codes on the document or separator pages.
- **Pattern matching:** Identifies documents based on logos, images or other graphics.

In addition to the standard capabilities available in capture software products, Datacap offers the following additional, advanced capability:

- **Fingerprint matching:** Identifies documents based on the geometry of page image by comparing the image against known page “template” images.

There is no one identification method that works for all documents. The capture software must permit multiple methods of identification and multiple layers of identification. For example, an initial step to identify a document might be to look for a logo or image in a particular location on the page. If the logo is not found, then the software might look for a bar code. If a bar code is not found, you might then want the software read the entire page performing optical character recognition (OCR) and search for keywords.

But what if your document is primarily unstructured text, with no logos, bar codes, or other clearly defining characteristics? Unstructured text documents might contain many of the same words, rendering text matching ineffective. How do you identify documents that fail classification by the standard methods? How can you automate document identification and avoid the pitfalls of manual sorting?

Wordfire Classify to the Rescue

Wordfire Classify is a component of the Datacap's Taskmaster suite of software for document capture and forms processing. It extends Taskmaster's traditional document identification capabilities with a unique, self-learning, text-analytics engine that delivers true human-like content analysis for the document identification process. Wordfire can accurately identify any unstructured, text intensive document, reducing or eliminating the expense of manual pre-sorting, insertion of separator pages, or adding barcodes prior to scanning. This helps speed document processing and enables you to reduce or re-allocate resources to other activities such as customer service.

Using technology developed for the intelligence community, Wordfire goes a step beyond traditional OCR/ICR, image analysis, or keyword/phrase identification. It reads documents looking for patterns, concepts, and associations, and stores the results mathematically. It can quickly learn an range of document types, and then will accurately identify newly scanned pages, eliminating manual classification. And it works under-the-covers – with no keyword databases to setup and maintain and no marking of locations or zones on the page to search.

Using Wordfire, you can reduce or eliminate the expense of manual pre-sorting, insertion of separator pages, or adding barcodes.

Provides Human-like Cognitive Recognition

Wordfire is the closest document identification solution to human cognitive recognition. And its learning is cumulative – what Wordfire knows about a given category of documents is the sum of all the example documents in that category. Wordfire is a very effective identification solution for textual documents because:

- It deals with concepts and documents in their entirety. Unlike solutions based on linguistic approaches, Wordfire is independent of sentence structure – increasing classification accuracy.
- It is virtually impervious to OCR and input errors. OCR errors will “throw off” solutions based on natural language and linguistic concepts – this is not true with techniques of Wordfire because it learns for real-world documents, and the real-world OCR results from reading those documents.
- It is self-training. Techniques that rely on physical positioning and textual “context” need to be guided through the machine learning process, i.e. be told what to learn. Wordfire ingests the whole document and learns solely based on the concepts contained in the document.



Exception Processing

Wordfire excels at handling the “exception” documents, those that cannot be identified by the traditional methods of character, barcode, pattern or fingerprint matching. These documents may be “kicked-out” by your current workflow rules, and end up being manually identified wasting time, resources and money. *Wordfire* provides the opportunity to automatically identify these exceptions, eliminating manual classification. Users of Datacap Taskmaster can keep existing Taskmaster workflows and add *Wordfire* as a final method of identification before resorting to manual classification.

Unstructured Documents

Wordfire excels at identifying unstructured, word-intensive documents. Documents that are pages of text, such as contracts, depositions, and affidavits, often require review by a person in order to understand the content so they can then be identified and processed. The text analytics of *Wordfire* enable it to “understand” these word intensive documents and properly classify them. It reads documents – looking for patterns, concepts, and associations and stores the results mathematically. Other methods of identification do not work with this class of document, forcing the expense and delay of manual processing.

Distinguishing Features

Wordfire is a unique document identification solution that delivers a quick ROI:

Easy to start and manage– *Wordfire* trains itself using your documents and automatically builds the conceptual associations between the content and the classification categories. No keyword database is required. *Wordfire* then accurately classifies your documents without intervention.

Rules-based – *Wordfire* is rules-based so that it is easily configured, adapted, and integrated with other systems using the Rulerunner Service from Datacap. Datacap rules allow the user to control the processing very precisely, calling *Wordfire* for all documents, or just documents that need to additional classification.

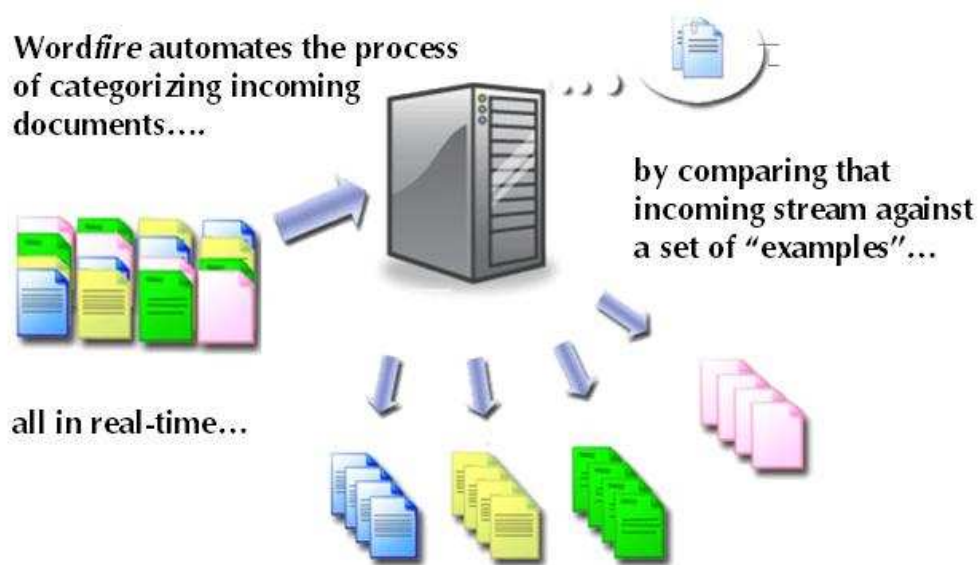
Independent of sentence structure – *Wordfire* deals with concepts and documents in their entirety unlike other solutions based on linguistic approaches, increasing classification accuracy.

Virtually impervious to OCR and input errors – The mathematical techniques of *Wordfire* greatly reduce the effect of OCR or spelling errors that “throw off” solutions based on natural language and linguistic concepts.

How it Works

Wordfire Classify is easy to install and requires virtually no setup and no ongoing maintenance. It works with Datacap Taskmaster version 7.1 or higher. The Wordfire software may be installed on the Taskmaster server, a separate server or run on the client machine. It requires a database for storage. Supported databases include SQLServer, Oracle 9i, Oracle 8 and MySQL.

Applications using Wordfire are created with Datacap Studio, Datacap's user-friendly design and test tool. Standard Taskmaster Actions (the building block of rules) perform the OCR operation on the image documents awaiting identification/classification by Wordfire. The OCR results are passed to Wordfire for comparison against the library of examples to identify the document type. The document type is returned to the Taskmaster application and processing of the documents continues.

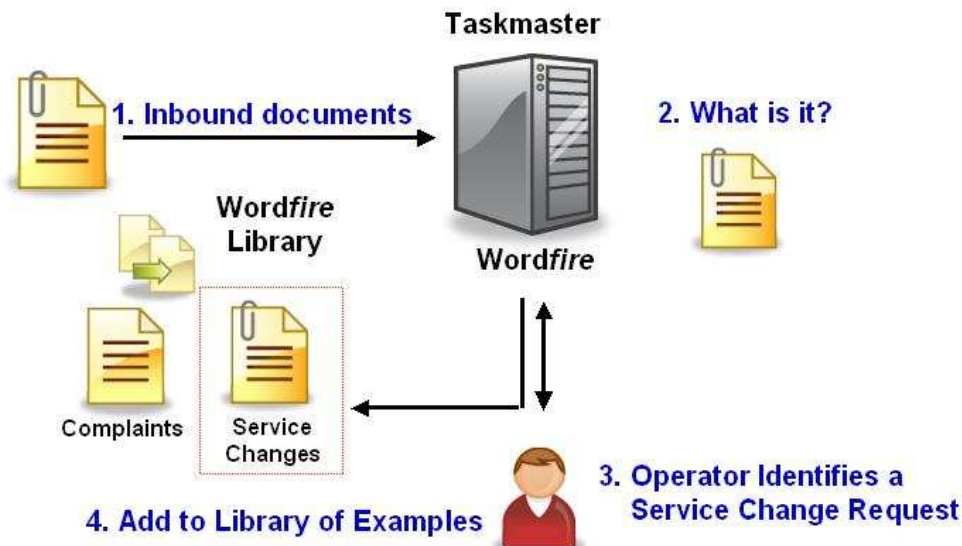


The Learning Process

Wordfire is self-training, just start using Wordfire and it begins to learn about your documents as you go along. When a new type of document is encountered, Wordfire presents it to an operator who then chooses the document type from a pull down list. This new document is then added to the library of examples, increasing Wordfire’s understanding of your documents. “

A diagram of the process is below.

1. You begin using Taskmaster and Wordfire to process your documents. (Taskmaster can initiate the scanning or the documents may already exist as image files.)
2. Since you have just started using Wordfire, the document types are unknown.
3. Taskmaster displays the document to an operator and the operator identifies the document type from a pull-down list of available types.
4. Taskmaster then performs OCR on the document to extract the text and sends the text and document type to Wordfire. This builds up the library of document examples, which will be used to automatically identify, or classify, other incoming documents.

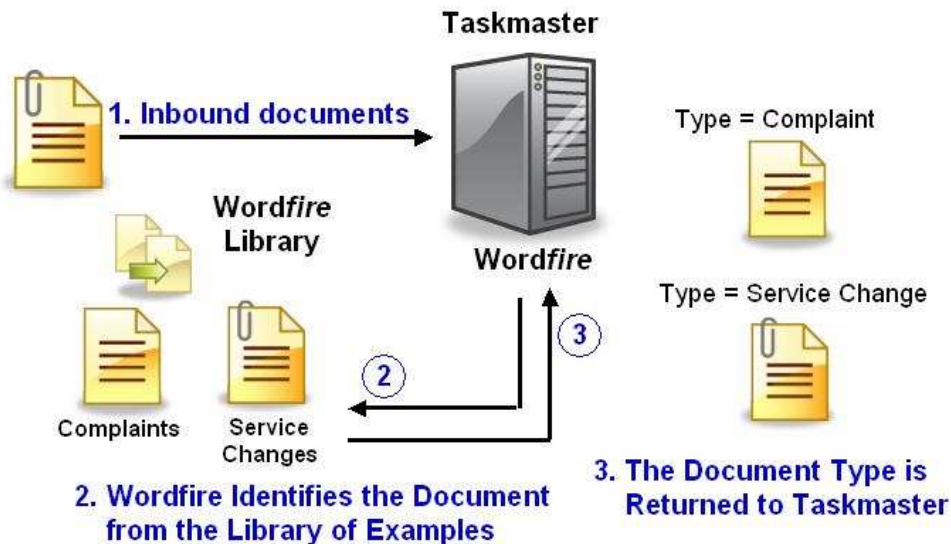


The number of example documents needed depends on the number of different document types. As a rule of thumb, a dozen to two dozen sample documents can serve as the basis of accurate document classification with Wordfire. In one test of 1000 mortgage documents, 50 examples of eight distinct document types proved sufficient for correct identification the complete set of documents.

Production Use

As the library of examples grows, *Wordfire* begins to recognize more and more incoming documents. Eventually it has enough examples and the process of storing new examples can be turned off. *Wordfire* can now automatically identify all your document types. Using *Wordfire* from the Taskmaster environment requires the use of just a few Taskmaster *Actions* (the building block of rules). *Wordfire* compares incoming documents to the examples in the library, identifying the document type based on a probability of a match, a value between .0 (lowest confidence) and 1.0 (highest confidence).

In the diagram below, notice the documents entering into a Taskmaster application for identification using *Wordfire*. Taskmaster performs full-page OCR on the document and passes the results to *Wordfire* for analysis. *Wordfire* analyzes the document text and compares it to the example documents in the library looking for patterns, concepts and associations. If a match is found the document type is provided to Taskmaster. In this example *Wordfire* has identified a complaint letter and a service change order. No person needs to be involved in document identification and processing.



Summary

Wordfire accurately identifies unstructured and text-intensive documents, reducing or eliminating the expense of manual pre-sorting, insertion of separator pages or adding barcodes prior to scanning. Like a person, it reads documents – looking for patterns, concepts and associations and stores the results mathematically. Soon it will understand your entire range of document types and will accurately sort them, eliminating manual identification. *Wordfire* speeds document processing and enables you to reduce or re-allocate resources to other activities such as customer service.

Benefits

Easy to get started, *Wordfire* learns from your documents without creating and maintaining a keyword dictionary. It easily integrated with any scanning or capture solution and seamlessly integrates data into any enterprise applications including ERP and document management or archiving systems.

Wordfire provides the following benefits:

- Eliminates pre-scan manual sorting and document separating
- Enables automatic processing of mixed document batches
- Processes “noisy” OCR documents without operator intervention
- Increases customer satisfaction with faster and more accurate document processing
- Improves document processing efficiency
- Reduces error resolution time and costs

About Datacap

Since 1988, Datacap Inc. has provided award-winning document capture and forms processing software solutions to organizations worldwide. Datacap Taskmaster software accurately transforms paper into information, increasing efficiency while reducing costs and document cycle time. A client/server, rules-based capture workflow platform, Taskmaster provides the industry's most adaptable solutions for both document image indexing and forms processing. Taskmaster also enables scanning and indexing from a browser and integrates with all leading document management solutions, databases and ERP systems. Awarded AIIM "Best of Show" for Forms Processing, Rulerunner Service provides the only complete capture solution for service-oriented architectures (SOA). For additional information, visit www.datacap.com.