CRAWFORD TECHNOLOGIES

WHITE PAPER

# Document Data Mining

## An Untapped Vein of Value

wwww.crawfordtech.com sales@crawfordtech.com

## Introduction

Data mining is an exotic-sounding term that conjures images of forensic computer analysts plucking nuggets of information from vast stores of digital data and using it to identify trends and connected behaviors. While mining operations can be complex and mysterious, lots of applications for extracted data do not involve searching for hidden patterns and correlations. A bountiful source for useful data, and one often overlooked, are everyday business documents.

**The best source of data necessary for a wide variety of uses is often overlooked. Extracting data from documents may be the best approach.**

Documents already contain information gathered from various sources across the enterprise. This data can enable document operations to be more efficient and accurate, or take advantage of new technology. It can help businesses understand their customers better, build more relevant communications, and improve the customer experience. Additionally, data mined from archived documents can help companies develop marketing strategies or follow regulatory directives.

Extracted document data can populate inserter files, create indexes for archived pages, or convert messages into alternative forms such as creating accessible communications. It can also be used to drive transpromo messaging, re-sequence print files, or combine mail pieces as part of a householding strategy.

Companies wishing to exploit the advantages of white paper workflows will almost certainly be interested in extracting data from print files. Controlling the integrity of mail pieces generated from several jobs after merging them for print production is imperative. Barcodes constructed with data extracted from the original pages ensure every page is accounted for by the finishing equipment. Extracted data can also make automated reprints possible, in case a document is damaged.

## A Competitive Advantage for Service Providers

Print/mail service providers frequently do not have access to the raw data used to compose the documents they are charged with printing and distributing. Their only source of data is contained in the print file. Even basic operations such as extracting address data for postal processing or determining page counts relies on their ability to inspect, remove, and sometimes replace data that exists on the printed pages.

Data mining tools can allow print service providers to streamline their own processing workflows and add value for their customers. A shop producing investment account statements, for instance, might segment high-balance accounts from the rest of the print run. They could add reports, print statements on higher quality paper, or assemble the pages into portfolio binders instead of folding and inserting statements into window envelopes. Offering more functionality fueled by data mining distinguishes print service providers from their competitors and allows them to charge higher rates.

## Combine with Outside Data

Sometimes, extracted data can be joined with other information beyond the source documents to compose entirely new documents which may be delivered on paper or digitally. A company could, for example, combine credit card transaction data with demographic customer information to create relevant offers corresponding to a customer's purchases, income level, or age. The offers may be delivered as embedded ads on the credit card bill, as calls to action in digital bills, or even used to selectively insert pre-printed marketing material into envelopes.

In another example, data about individual customers may reside in a CRM database, but not in the documents themselves. It is necessary to extract match keys, such as account numbers, from the print file to access more information from the CRM system. This is important when creating accessible documents as customers may demand different document formats, depending on their abilities and resources. Customers may need large print, braille, or auditory delivery forms of the documents. Companies are obliged to satisfy reasonable accommodation requests, which are recorded in the CRM database. Connecting the document data to external data sources is an efficient way to comply with the law.

## Why Get Data from Documents?

It may seem odd, looking to documents as a source of data, but there are advantages to accessing this readily available information as an alternative to acquiring the data from the original sources. Accessing records stored in far flung databases scattered throughout the enterprise requires the services of IT specialists, secure connectivity, and data structure knowledge.

Obtaining funding and support for such initiatives can be a lengthy and uncertain process. Most times, pulling needed information right from the documents is one of the few ways projects will get done within a reasonable time.

Document archives are static, whereas data warehouses usually contain only the most current information. Any project that requires information from the past is more likely to find accurate data stored in the pages that were created at the time.

As the only trusted source for this vital data, it's clearly essential that the information be complete and accurate.

Since the content within bills and statements often comes from many disparate systems, the collection of data is considered the official record of transactions that must be maintained to meet the organization's regulatory obligations regarding customer communications. As the only trusted source for this vital data, it's clearly essential that the information be complete and accurate.

A good example is legal discovery. Companies responding to litigation, audit, or regulatory inquiries can find themselves in a bind. Finding and extracting data stored in old versions of multiple systems (that may have changed over time) will require spending money on outside service providers and/or the attention of skilled individuals from the company's IT staff. Either way, compliance will be difficult, disruptive, and expensive.

Without a comprehensive data history, information collected from document archives can satisfy the legal mandates. What might have taken months to achieve when dealing with the raw data can be accomplished in weeks if the e-discovery data can be pulled from the document storehouse.

## Using the Right Tools

This doesn't mean extracting document data is easy. Documents are designed for consumption and interpretation by humans, not machines. Common items programmers use to identify data in databases such as tags or fixed file structures are missing from print files. Locating specific information locked in documents requires techniques such as data markers, offsets, rules, and verification. Documents frequently feature conditions that result in exceptions or unique circumstances, which must be handled as part of the data mining operation. Isolating the correct data involves combining physical page location, inspection, and logic. Without great tools to find and interpret the data, the process of extracting information from print files is tedious. It may even be impossible.

Mining data from documents is often performed by forms analysts, document designers, or business unit staff members. They are not computer programmers. Constructing the routines to find and extract the data must be done in a graphical manner, using a friendly interface. Highlighting data fields on a screen is infinitely easier than manually computing x and y coordinates or writing IF-THEN-ELSE logic by hand. The ability to create embedded, non-printing data elements within documents stored in archives is a powerful capability, and greatly simplifies data extraction when accessing the archive in the future.

The ability to create embedded, non-printing data elements within documents stored in archives is a powerful capability, and greatly simplifies data extraction when accessing the archive in the future.

## Resourceful Uses for Extracted Data

- Uses for extracted document data are nearly unlimited. Advances in printing technology and digital delivery channels have made it possible for document developers to make messages more personal, relevant, and effective than ever. Here are a few examples:
- Delivery address data from a bank statement can trigger maps or personalized driving directions for customers who live near a newly opened branch.
- Item purchase transaction details can generate a QR code leading to instructional videos that answer frequently asked questions about purchased products, cutting down on product returns and increasing customer satisfaction.
- Transactional information related to past payments can be used to suppress mailing a remittance envelope to customers who always pay online.
- Analysis of services listed on bills enables the generation of marketing messages for upgrades or additional services – and prevents the company from marketing for services to which customers are already subscribed.
- The ability to create embedded, non-printing data elements within documents stored in archives is a powerful capability, and greatly simplifies data extraction when accessing the archive in the future.
- Account, customer, and invoice information included on a bill are used to construct a printed Personal URL leading to a web page featuring personalized incentives for converting to paperless billing.

- Address information on transactional documents drives selective marketing messages for partner companies or special events near each customer's home. Some organizations sell ad space in bills. Precise targeting makes this feature more appealing to potential advertisers.
- Product purchase information extracted from documents can be used to trigger follow-up emails soliciting customer reviews and feedback.
- Mining for data from documents can be vastly more efficient than retrieving the same information from scattered databases. Document designers have already done the work of connecting data from various sources. Organizations will not have to rely on IT resources to obtain the data they need to improve operations, lower costs, respond to legal inquiries, or enhance customer communications.
- Document owners should not overlook the value hidden in items they already have. Archived documents often represent the most accurate representation of data as it existed during a certain period of time. Given the right tools to extract the data, companies will find their documents are a valuable source of easily-acquired information.

Mining for data from documents can be vastly more efficient than retrieving the same information from scattered databases. Document designers have already done the work of connecting data from various sources. Organizations will not have to rely on IT resources to obtain the data they need to improve operations, lower costs, respond to legal inquiries, or enhance customer communications.

Document owners should not overlook the value hidden in items they already have. Archived documents often represent the most accurate representation of data as it existed during a certain period of time. Given the right tools to extract the data, companies will find their documents are a valuable source of easily-acquired information.

## CrawfordTech Solutions

Crawford Technologies develops software and solutions to help enterprises optimize and improve the secure and accessible delivery, storage and presentment of their customer communications.

With over 1,800 customers on six continents, CrawfordTech solutions and know-how enable the largest banks, insurers, healthcare providers, utilities and print services companies to use their existing technologies, documents and data in new ways. We help them navigate the challenges in leveraging legacy applications in the platforms and applications of the future.

CrawfordTech's products, services and domain expertise reside at the nexus of content, data, and output management and are essential components of our customers' digital transformation, output management and document accessibility strategies.