Businesses are seeking productivity increases and launching cost reduction initiatives that require a new way of thinking about business documents and how they relate to business process and transaction-oriented data.

Most experts and analyst groups claim that only 20% of data is structured while unstructured is closer to 80%. Make sure you can utilize this important data resource.

**Takeaways**
- Understand important data types in forms and documents
- Know how to determine the right method to capture data in business process and transactional systems
- Learn strengths and weaknesses of three capture categories: Ad Hoc Image, Batch Image, and Batch Transaction
- Discover how to utilize technologies that can make a powerful difference in your business
What’s in this Paper?

- Introduction page 2
- How Unstructured, Document-based Data Can Be Classified page 2
- Structured and Unstructured Forms page 3
- Non-form Documents page 4
- The Critical Role of Capture Solutions page 5
- Importance of a Disciplined Approach to Capture page 6
- Recognition for Unstructured Forms and Documents page 7
- Conclusion page 9

Introduction

Increasingly businesses are seeking productivity increases and launching cost reduction initiatives that require a new way of thinking about business documents as they relate to business process and transaction-oriented data.

Historically, information either came in as electronic, structured data or it was forever “trapped” in its physical form. The generally accepted terms were “structured data” vs. “unstructured data”. Most experts and analyst groups claim that only 20% of data is structured (with structured data being that stored within databases) while unstructured is closer to 80% (e.g. non-relational database data stored within documents including forms, receipts, correspondence, and internal documents).

Within this unstructured domain, however, there are important nuances to the types of data and how they are represented that are critical when determining the right method to incorporate into process and transactional systems.

How Unstructured, Document-based Data (the other 80%) Can Be Classified

While most experts agree that 80% or more of business information is received, stored, or sent as unstructured non-relational data within documents, when we dig
further into document characteristics, there is actually a continuum of “structure” within this arena. Most typically, business documents can be grouped into two overall categories: Forms and “everything else”.

**Structured and Unstructured Forms**

Forms can encompass everything from applications for credit to magazine subscription cards. Merriam Webster defines a form as “...a printed or typed document with blank spaces for insertion of required or requested information. Example: tax forms.”

This definition leaves a lot of room for variability and it is precisely that variability that determines the amount of effort required to address document capture needs for any organization. To add more context to the discussion, it is helpful to categorize forms into one of the following three categories: Standardized Structured Forms, Variable Structured Forms, and Variable Forms.

**Standardized Structured Forms**

These are documents that are always structured in the same manner and provide data uniformly regardless of the business or government entity using the form. The focus on these forms tends to be very transactional and supportive of well-defined processes.

Typically standardized structured forms are designed and implemented by a standards or regulatory body for use within a certain industry or for a specific purpose. Public examples include health claims forms such the CMS 1500 used by Medicare/Medicaid. Private examples include the Automotive Industry Action Group’s trade specifications. Since these are very standardized document structures, in most cases, the physical standardized form is also accompanied by an electronic format typically using Electronic Data Interchange.

**Variable Structured Forms**

These are documents that, to a large degree, have the same uniform data, but the structure is different depending on the organization. The primary reason for the variability involved with this category of forms is that individual organizations can create and manage their own form layout.

Like their standardized cousins, these forms also are focused on collecting and delivering transactional data but there is often a challenge getting that data into processes and systems. A common example is a company invoice which almost
always includes the following important data: invoicing company, invoiced company, address, date, line items and amounts, total, and due date.

Unstructured Forms

These are documents that fulfill a common business need, but that contain both variable structure and variable data. Data within a specific class of variable form can be the same but still have differences. The data might not always support a transactional process – typical uses include informational-only and/or record-keeping. This document type is often referred to as an “unstructured form” but in reality there is still a structure that can be defined. The challenge is often with gathering a comprehensive understanding of all potential document structures and data types.

Examples of these documents types include surveys, medical records where information can be the same but contained within highly variable structures, and membership forms where required information often differs depending on the organization (free membership or paid, health requirements, etc.).

While all the above form types share common characteristics such as a defined structure or common data types, standardization and consistency across organizations is the key variable that determines the level of complexity when incorporating a specific form type within a document capture strategy.

Non-Form Documents

Outside of business forms designed to collect data, there are a number of documents that are considered essential to business operations but that do not have a pre-determined form. These are true unstructured documents, and this category includes customer correspondence via letter or email, doctor’s notes, manuscripts, court records, video, and images.

Given the largely undefined and unconstrained use of documents within an organization, it is not surprising that the previously mentioned yardstick of “80% or greater” is a common estimate of unstructured documents compared to all business documents.

These documents have definite value within the organization and could include information crucial to a transaction or process but in a less defined way. Most provide context for a transaction – for example, a customer complaint. With this document type, the challenge is to first identify the type of document, and then to extract the most important data.
The Critical Role of Capture Solutions

There are a wide variety of solutions that enable organizations to capture business-related documents and incorporate them into business processes. While features and capability can vary based upon the industries and markets served, the Association for Information and Image Management (AIIM) has defined capture to consist of three major categories: Ad Hoc Image, Batch Image, and Batch Transaction. Each has its own strengths and weaknesses when it comes to addressing document processing needs. We will briefly explore each category.

Ad-hoc Image

Ad-hoc image solutions typically involve capture from desktops or multifunction printers that do basic document processing and some indexing. These solutions are document-centric in that the document itself is used as a key asset within a business process or transaction. As such, most solutions will typically involve storage and often retrieval as a key capability to enable an organization to easily find the needed information. Ad-hoc image capture is a core capability of many solutions and, regardless of capture maturity within an organization, will always serve a specific need.

For forms capture, ad-hoc image processing can do the job but is often a very time-consuming solution due to the need to manually index each document as it gets captured. A few solutions have addressed this issue with basic batch class capabilities that allow for some automated data extraction via optical character recognition (OCR) but there is still a lot of manual intervention required if data extraction needs go beyond basic document description indexing.

Interestingly enough, ad-hoc image capture solutions are often used to assist with indexing of unstructured documents using both full-text OCR and human-based processes to examine and then identify the document type.

Batch Image

Batch image solutions are document-centric in nature but increase the scale and volume by using networked department-level scanners. Solutions within this category include the ability to scan multiple documents as batches that share common business/department purposes. They also support higher-volume document indexing through the deployment of an “indexing client” that allows users to view documents and enter key data that describes the document.
Batch image capture originated the idea of forms processing. These solutions introduced the concept of batch classes that use pre-defined form templates to automate both the separation of document batches into logical document types, and the application of template rules to automate the extraction of data on the form based upon data location.

Organizations that are looking to expand beyond basic image capture typically start with batch image capture as it provides an organization with a very effective way to not only reduce document storage and retrieval costs, but also the ability to begin using documents to trigger and assist with common business processes.

**Batch Transaction**

The third capture solution category as defined by AIIM includes software specifically designed for high-volume data extraction from physical documents. In this category, the document no longer is the key asset within a business process or transaction; rather, batch transaction essentially enables an organization to turn the document into just another “data transport protocol” where the data itself is extracted, validated, and sent to other transaction or process applications. Tasks such as indexing play an important supportive role to ensure data extractions are accurate and complete and often an exception handling event only.

Because automation is a key aspect of batch transaction capture solutions, forms capability is an essential area that requires the ability to operate in an “unassisted” setting. This means that wherever possible, document types must be automatically identified; the required data on the document must be extracted, validated, and then transmitted to the constituent system(s). Automation across all document types is the focus of the last several years as software vendors seek to take their solutions to the next level.

Additionally, it is becoming more common to have a tighter integration between transactional capture and business process platforms to create a more coordinated flow-through of transactional data within business workflow rules.

**Importance of a Disciplined Approach to Capture**

While each category of capture is more sophisticated than its predecessor, it is not uncommon to have all three capture types operating within an organization. More often than not, however, ad-hoc image is adopted first, followed by batch image, and lastly batch transaction. The key is to understand the type of capture necessary to support and accelerate transactional and business process needs. As business needs
become more sophisticated and internal capabilities expand, organizations will look to capture solutions to support the broadest range of documents, with more automation, and more tightly coupled to business processes.

Recognition for Unstructured Forms and Documents

The largely uncharted territory within the capture realm is the incorporation of unstructured forms and documents within business processes and transactions. As noted in previous sections, the desire for comprehensive automation combined with the highly variable nature of these document types turns a seemingly simple problem into a very complex set of problems. It is no longer practical to have employees manage document sorting, identification, and data extraction. Rather, solutions need to automatically sort incoming documents according to characteristics of the documents, apply appropriate business rules to extract and validate data, handle exceptions quick and efficiently, and then send the finalized data to one or more systems simultaneously.

Pattern Recognition for Data Extraction and Document Identification

Solutions emerging in the market that tackle the unstructured form and document needs of organizations often employ similar tactics to automate document capture. The underlying technologies can be dramatically different, but the concepts are similar.

The first capability is pattern matching applied to key words or symbols. This technique was first used to help organizations manage the large variability of two very common forms: invoices and purchase orders. While the data within these types of documents is very uniform, the placement of that data is far from uniform.

Without some allowance for variability, an organization using tried-and-true methods of pre-defined form templates common within batch image capture would face a daunting task of first identifying all the types of variations of form structure, implementing specialized form definitions (that could number into the hundreds) and then maintaining them and managing exceptions when either new forms are added or existing form structures change.

Pattern matching allows organizations to apply rules to the data itself and not to the nature of the form. This means that data types can be identified either with the common labels used to identify them on the form or through common data patterns that can distinguish one data type from another (e.g. social security numbers,
telephone numbers, etc.). When a document is captured, it goes through a process where a portion of the document or the full-page is scanned, the extracted data is sent to a decision engine that matches output with pre-defined keywords or patterns used to determine what data fields are to be extracted.

The above process assumes the document type is known (e.g. “invoice” or “purchase order”). Therefore, batches are still pre-sorted as much as possible into batch classes that can apply the appropriate data extraction rules.

With more advanced solutions, pattern matching goes even further to help automate the classification of the document. In these solutions, the goal again is to automate as many steps as possible. As with data extraction, document identification uses keyword or data pattern matching. Some solutions go even further by including the ability to recognize a form by corporate logos or some other image on the document.

**Semantic Recognition**

The next evolutionary step with capture for unstructured forms and documents is the ability to extract meaning from documents that lack any definable structure. With any type of form, it is possible to apply some moderately sophisticated rules to identify presence of needed data, where that data resides, then extract the data and send it on its way.

However, when the documents truly lack structure, it is impossible to go the same route. Required data may reside in one or more sentences together which provide the actual meaning. Take for example, the simple action of responding to a request to cancel an insurance policy. Depending on the process, this can be easy or incredibly difficult. Using a form, this type of action can be handled by a simple checkbox to cancel or modify insurance. But when it comes to recognizing the phrase “I would like to cancel my insurance policy” in the form, it can be very difficult. Most organizations handle written requests that lack structure as an exception which requires a person to evaluate the document and categorize it. But exception handling is very costly, especially when it can be automated.

New solutions are combining document recognition technology with semantic text analysis to categorize and extract meaning from even the most unstructured documents including hand-written letters, doctor’s notes, and legal documents. To date, text analysis is more common in search engines where the user does not understand how documents are classified or in some cases, what the right search terms are. In response, the search engine uses special indexing to attempt an understanding of the meaning and quality of the information within the document.
(as best it can). Thus, when a user enters a search for a topic, the results would ideally yield meaningful documents.

Applying text analysis to documents to automate the capture process is a relatively new field. In this application, the text analysis is accomplished at the time of capture to help determine the type of document and the nature of the underlying content. Assuming the document and the relevant data can be identified, business rules can be applied to pull (or sometimes derive) necessary data that is actionable and can be forwarded-on to other process or transactional systems.

Because an estimated 80% of business documents are unstructured, the use of text analysis can provide major improvements to transactional capture applications such as digital mailroom initiatives, and comprehensive medical records solutions. Solutions that allow for transactional capture of all document types with little to no human intervention, and little set-up required, represent a significant leap forward in effective capture within an organization.

**Conclusion**

While document capture has been around for several decades, the evolution of business processes and transactions from very manual to highly efficient and automated has had the most significant impact to the required capability of capture solutions available within the market.

When examining capture from the point of view of both a business process or transaction and the type of documents typical within the process or transaction, the appropriate solution becomes clearer.

<table>
<thead>
<tr>
<th>Structured Form</th>
<th>Ad Hoc Image</th>
<th>Batch Image and Zone Recognition</th>
<th>Keyword and Pattern Recognition</th>
<th>Semantic Recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ad Hoc Image</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Batch Image</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zone Recognition</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Keyword and Pattern Recognition</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Semantic Recognition</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable Structured Form</th>
<th>Ad Hoc Image</th>
<th>Batch Image and Zone Recognition</th>
<th>Keyword and Pattern Recognition</th>
<th>Semantic Recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ad Hoc Image</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Batch Image and Zone</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recognition</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Keyword and Pattern</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recognition</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Semantic Recognition</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Unstructured Form</th>
<th>Ad Hoc Image</th>
<th>Batch Image and Zone Recognition</th>
<th>Keyword and Pattern Recognition</th>
<th>Semantic Recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ad Hoc Image</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Batch Image and Zone</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recognition</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Keyword and Pattern</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recognition</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Semantic Recognition</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Unstructured Document</th>
<th>Ad Hoc Image</th>
<th>Batch Image and Zone Recognition</th>
<th>Keyword and Pattern Recognition</th>
<th>Semantic Recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ad Hoc Image</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Batch Image and Zone</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recognition</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Keyword and Pattern</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recognition</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Semantic Recognition</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Semantic Capture is still in its relative infancy with most organizations still looking to incorporate document capture of forms-based documents into their processes and transactions. As real deployments continue to prove a positive return and solutions continue to mature making production use simpler to configure and maintain, semantic capture will become the next logical step within organizations looking to apply a consistent and holistic capture strategy.

About Parascript

Parascript delivers the world’s best Image Recognition Engine. Employing patented digital image analysis, handwriting analysis and advanced pattern recognition technologies, Parascript improves important business operations in areas like forms processing, medical imaging, postal automation, signature verification and fraud detection. The powerful Parascript engine processes over 100 billion imaged documents per year. Fortune 500 companies, postal operators, major government, and financial institutions rely on Parascript products, including the U.S. Postal Service, Bell and Howell, LLC, Fiserv, Elsag, IBM, Lockheed Martin, NCR, Siemens, Xerox and Burroughs. Visit Parascript online at http://www.parascript.com.