Intelligent Document Understanding

Transforming disparate data into valuable knowledge

A white paper from
Kodak Document Imaging

Kodak

Introduction – dealing with a whirlwind of sources, data and formats

There are individuals, no doubt, who fondly recall the days of simplified document management. There were forms, correspondence, receipts, purchase orders and so forth, but they were all quite straightforward and arrived as pieces of paper.

These documents arrived at a company, were scanned and entered into a Content Management System, then accessed by or sent to the appropriate individuals, departments, repositories or other appropriate destinations.

There was no need to worry about e-mail correspondence, **Tweets, Facebook** posts, images from smartphones, SMS/text messages, or multiple other information sources that are now part of our wired, social media world.

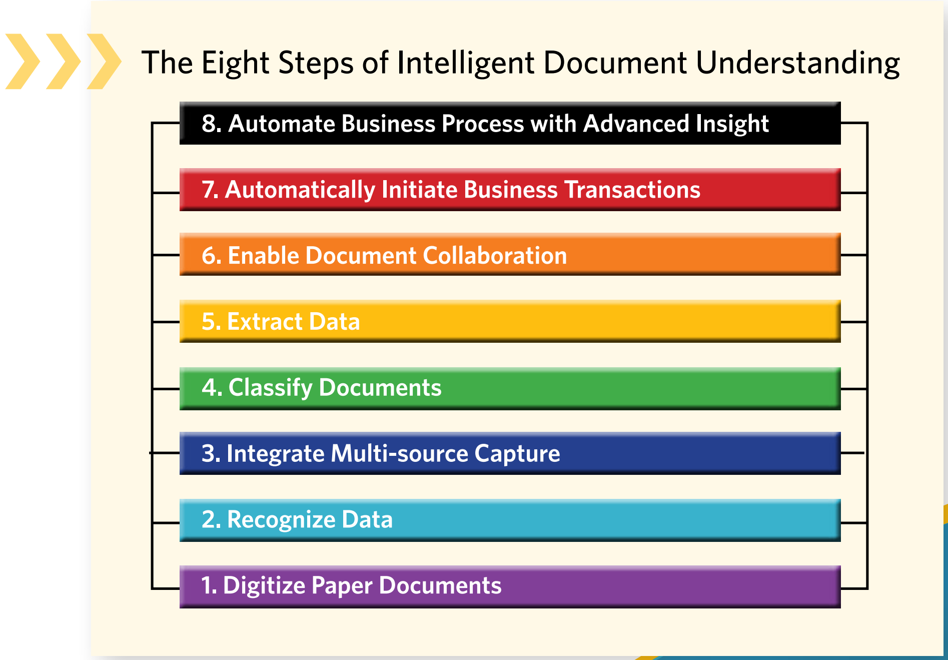
Of course, the evolution of Electronic Content Management Systems address a portion of these challenges and help to automate and better govern information capture, management and storage. But many aspects – such as multi-channel capture – remain cumbersome, due to the volume and complexity of formats and limitations of technology.

Welcome to multi-source, multi-channel information insight

Today, there is innovative technology available to better aggregate disparate data and make it intelligently useful for the enterprise and automate business processes. Advanced capture, Intelligent Document Recognition (IDR), classification and data extraction, multi-channel capture and other emerging solutions that go beyond paper to address that whirlwind of sources, data and formats.

This white paper from Kodak offers an overview of:

- Leading approaches to these business challenges
- Technologies that are available and evolving
- Practical, easy-to-understand real-world examples
- Key considerations and questions to ask technology solution providers as you work to enhance business processes and investigate potential systems



The Eight Steps of Intelligent Document Understanding

8. Automate Business Process with Advanced Insight
7. Automatically Initiate Business Transactions
6. Enable Document Collaboration
5. Extract Data
4. Classify Documents
3. Integrate Multi-source Capture
2. Recognize Data
1. Digitize Paper Documents

LET'S LEARN MORE ABOUT IT › › ›

How Intelligent Document Understanding can help drive your enterprise

Defining the data-to-knowledge continuum

Intelligence increases as you progress from data to information to knowledge. To illustrate this concept, we'd like to introduce you to Bob the robot. He will help us demonstrate the data-to-knowledge continuum.

1) Bob can read and speak, but doesn't know math. We ask him to multiply 13 x 9.

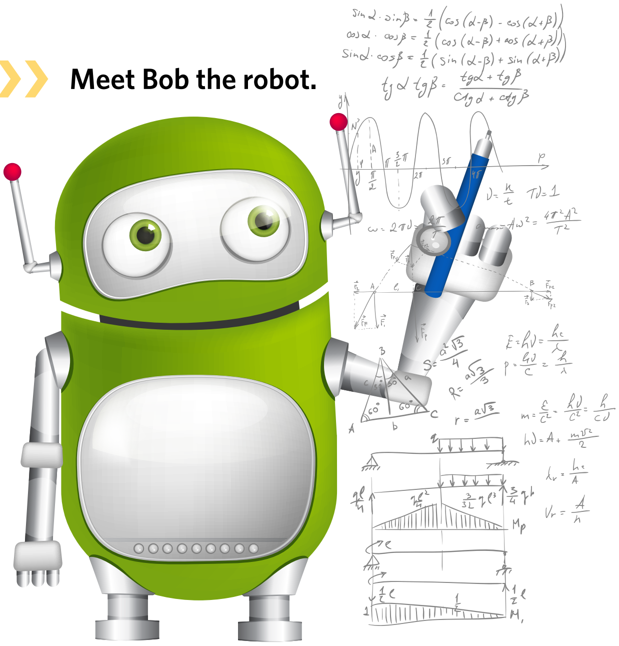
2) Bob sees four characters: 1, 3, x and 9 but does not know what these mean. This is **data** – raw and without meaning.

3) Now Bob learns the multiplication tables up to 12 x 12 = 144. He still does not know the answer, but knows this is multiplication: 13 times 9 equals something. Bob now understands the **meaning.** This is the definition of **information.**

4) If we ask Bob what 2 x 2 equals, he knows the answer is 4, because he's learned this as part of his multiplication tables. This is **knowledge** – and in this case rules-based knowledge and basic knowledge, as it's a quite basic problem.

5) Next Bob learns addition and subtraction, but is still limited to a problem set of 12. He now has more analytical skills and can compute the problem, determining the answer of (12 x 9) + (1 x 9) =117. This demonstrates **complex rules knowledge:** still rules-based, but more complex. The entire field of arithmetic is rules-based.

❯❯❯ **Meet Bob the robot.**



6) What if we ask Bob what country has won the most World Cups? He doesn't know and his knowledge of arithmetic and application of complex rules are no help. Rules-based systems are limited and a question like this can't be scripted.

Bob needs **semantic/contextual understanding,** such as **Siri** Software provides to iPhone users, to understand the question and comprehend the meaning of "what country." He also requires **Artificial Intelligence** (AI) to categorize the problem and grasp that "country," "won" and "World Cups" are key words and terms.

So we connect Bob to the Internet and he uses a **Google** search to find the answer. He then calls on AI once again to assess results and determines the answer to be "Brazil."

This demonstrates the field of Artificial Intelligence and Machine Learning (a sub-field of AI) where an AI engine is trained, creates connections from the learning set, and answers more complex questions when it does not have rules to find the answer. If you have seen the IBM Watson technology, you have experienced a super Bob!

IN SUMMARY ❯❯❯

Data is raw information with no meaning. **Information** provides meaning but no solution to a problem. Rules-based knowledge begins to offer solutions to problems that can be scripted. AI-based knowledge can solve significantly larger problems that cannot be scripted.

Real-world situations ideal for Intelligent Document Understanding

Let's meet three human, non-robotic individuals who will also aid us in discussing a practical application of Intelligent Document Understanding in the real world.

First, Larissa is **a knowledge worker** in the insurance industry, employed as a claims processing agent. Next is Jim, a customer of the insurance company. His car was damaged when a neighbor's tree fell on it during a recent hurricane. And finally Dave, an insurance company agent in Jim's hometown.

The information process

The information about damage to Jim's car arrives at the insurance company, more precisely at Dave-the-agent's office and the main office from different channels (picture, electronic form, paper documents, etc.). It takes a good deal of time, many phone calls and repeated attempts in order for Larissa to collect all these information elements that she needs.

Once aggregated, Larissa reads each, recognizes data types and elements, and enters key data into the insurance company's document management system, associating a name (such as Jim, a customer with a policy) with a client field, a claim number, a damage amount, and so forth. She enters information (metadata) into the system.

Larissa then uses the insurance system to obtain details about Jim's policy and determine claim eligibility, based on rules she has learned, thus using her knowledge. With the information provided by the documents and insurance system, she can make a decision, process Jim's claim, and send him a check.

Larissa, her management team, and customer Jim would all appreciate if this process could be sped up and became more transparent throughout, in order to answer questions in a timely, accurate manner. The eight steps we will outline next will demonstrate that these goals are achievable.

Now, before we embark on a quick trip through the Eight Steps of Intelligent Document Understanding, we should note that – depending on individualized organization situations – the order of these steps may occasionally vary.

STEP 1

The Eight Steps of Intelligent Document Understanding

Digitize paper documents

Paper documents need to be digitized to make them ready for data recognition and machine readability in Step 2.

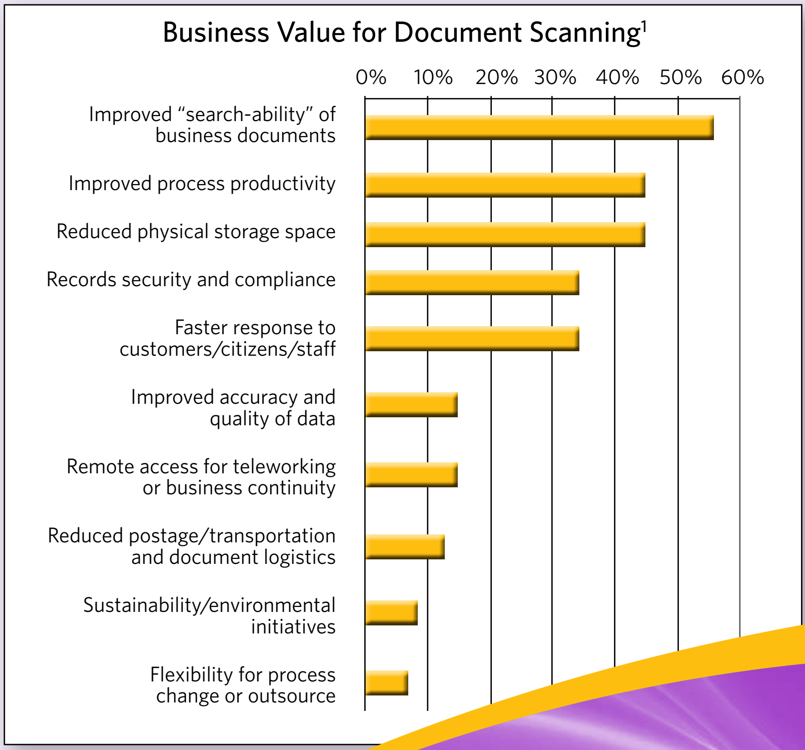
While you may hope that paper is going away, an AIIM white paper from 2012 states that the paperless office is far more dream than reality.[1] While paper document levels are decreasing somewhat, companies handling less paper are only slightly more prevalent than companies experiencing greater paper volume, according to the same AIIM report.

Real-world update: 45% of scanned paper documents are actually "born digital," as PDF files, faxes, Web forms, **Microsoft Office** documents, etc.[1]

Scanner
or
or

The business case for document scanning

Before advancing to Step 2, consider the key benefits of document scanning. The survey question: what would you say are the biggest drivers for scanning and data capture in your organization? "Search-ability" and "share-ability" are often cited as key business values for document scanning, as demonstrated in the AIIM survey results shown in the chart to the right.



Jim (the insurance customer) had his car repaired and takes the invoice to his insurance agent Dave. However, this is a key document that Larissa needs in order to work on the claim. If Jim mails the document to the head office this may take a couple of days, not even considering the risk of it getting lost. The ideal approach would be for Dave to scan the document right away and transmit the digital image to Larissa.

[1]Reference: AIIM, The Paper Free Office, 2012

KEY CONSIDERATION ❯❯❯

Evaluate your processes to determine opportunities for digitizing documents at the earliest point of entry. Depending on paper/document volume and logistical realities, a workgroup or personal scanner may be best suited for distributed scanning.

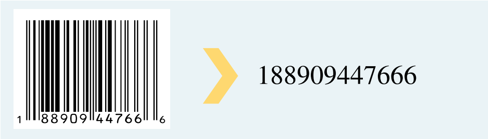
STEP 2 Data recognition

Having scanned and digitized the invoice in Step 1, this document has been converted into an image. In order to make it valuable for Larissa, we need to recognize the key relevant data elements. Helping Larissa to automate this recognition will result in significant time savings.

This is a good opportunity to take a look at the five common techniques for reading data from paper. Starting at the top, these techniques range from very accurate and fast … to less precise and more time-consuming.

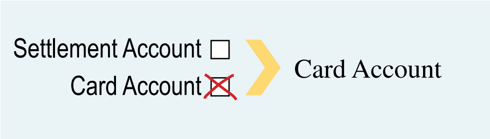
1. Barcode

Fast and accurate, but limited to a small number of document types as printing at the source is mandatory. In our example, the document can be identified as a claims report via an imprinted barcode.



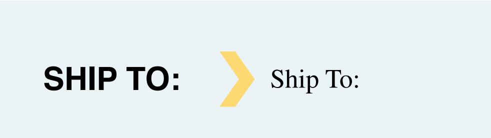
2. Object Mark Recognition (OMR)

OMR converts written marks in data fields, such as yes/no answers. Commonly used on surveys, applications, school tests and so forth. It's gradually being replaced by Webform capture. Barcode and OMR don't read an entire document, but extract key data from forms.



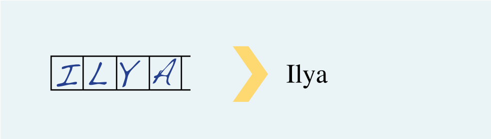
3. Optical Character Recognition (OCR)

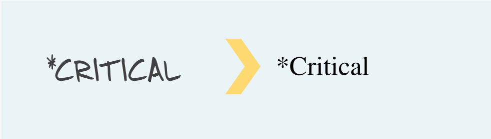
OCR recognizes printed text. It's now a mature technology and typically offers high accuracy on Latin-based characters. It is still in development with some Asian languages. OCR is a computing-intensive process that can slow PCs that don't have enough "horse power."



4. Intelligent Character Recognition (ICR)/ handprint recognition

ICR/handprint recognition is the least mature and by far the least accurate technology, unless severely constrained, such as in recognizing characters inside form boxes as shown; when using a limited vocabulary, such as 300 medical terms; with redundant information (reading checks); or when information can be matched in a database (reading postal addresses requires but a few characters to accurately determine a match in a postal database). Recognition of unconstrained freehand, cursive writing remains very challenging.





KEY CONSIDERATION ❯❯❯
Work with your solution provider to identify opportunities for applying enhanced automated data recognition or optimizing the current technologies applied.

Multi-source integration

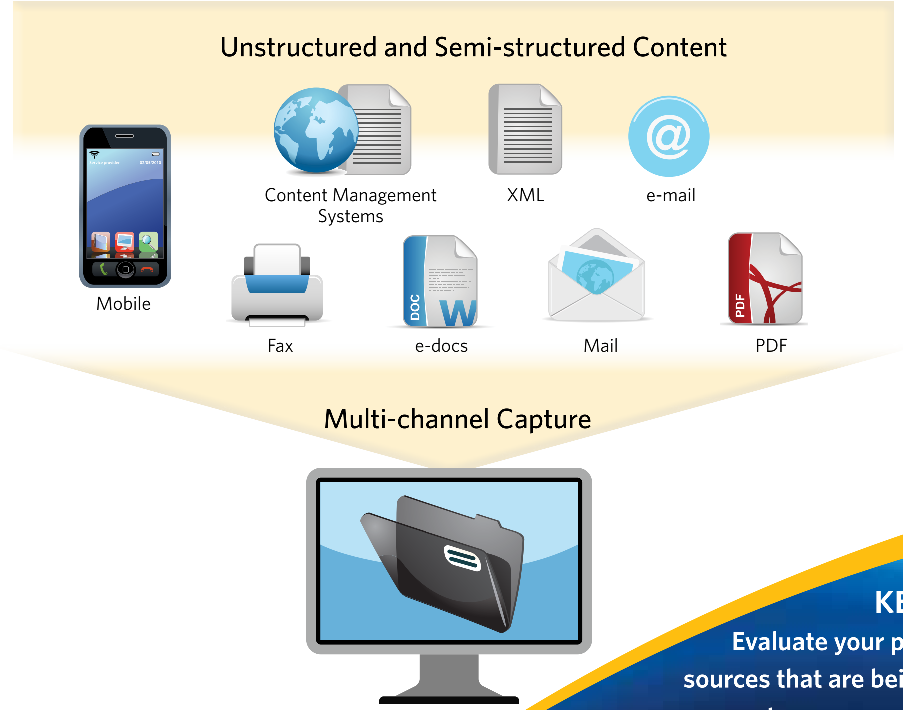
Today, expanding information sources (in both type and volume) carrying critical input for a business process – like our insurance claim example – create major headaches for companies.

Going back to our example: in addition to the invoice that Jim the customer handed to insurance agent Dave, he had already taken a photo of the damage from his smartphone and sent it to the insurance company. Unfortunately, Jim forgot to include his customer number. Jim then went online and completed a claim report electronically. Subsequently he called the insurance company and asked them to connect the photo he sent with his claim report.

In order to streamline this example, we have not even mentioned new channels like social media that should not be underestimated with regard to their volume and importance.

Today, few organizations can boast of an efficient, integrated multi-source capture process. And the sources of information continue to expand. If we revisit Larissa and our insurance example, customer Jim files his claim via Webform, calls his agent over the phone, transmits damage photos through his smartphone, mails the repair invoice and – depending on how he feels about the experience – may **Tweet** about it.

Currently, most of these input channels are still absorbing a good deal of Larissa's valuable time through manual processes. And those that are automated cover one input source only, such as paper-based documents or e-mail communication. Locating and connecting these input streams is a major challenge for Larissa in trying to process the claim. She must manually connect and aggregate all the separate inputs: Webforms, voice mails, photos and more into a case folder. This is a time-intensive process and she does not have much help from the existing capture systems that are primarily paper-based.


Unstructured and Semi-structured Content
Mobile
Content Management Systems
XML
e-mail
Fax
e-docs
Mail
PDF
Multi-channel Capture

KEY CONSIDERATION ›››

Evaluate your processes to identify information sources that are being handled manually or through separate processes where integration could lead to significant business advantages. A key selection criterion for your solution should be the capability to integrate not only scanned paper, faxes, and e-mail attachments, as well as e-mail body, text, electronic documents and social media postings.

STEP 3

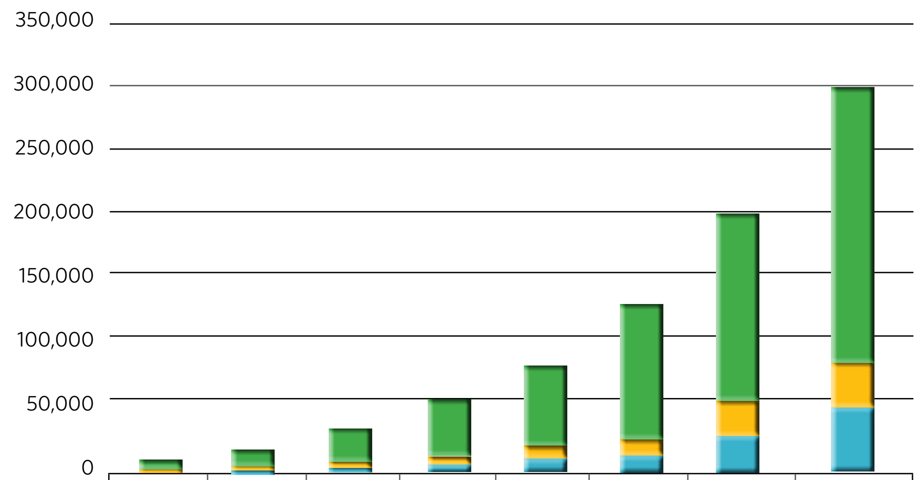
The Eight Steps of Intelligent Document Understanding

Multi-source integration, cont.

The growth and chaos of unstructured information

The 2011 AIIM ECM *State of the Industry* report asked how “well managed” various types of information were in organizations. Not surprisingly, paper was best managed, but content such as **Tweets,** instant messages, blog posts, e-mail attachments, voice/phone call records, e-mails, and photo images were viewed as in a predominantly chaotic or judged somewhat unmanaged. This presents both a problem and an opportunity to enhance business processes and collaboration for knowledge workers.

Total Archived Capacity, by Content Type, Worldwide, 2008-2015 (Petabytes)



	2008	2009	2010	2011	2012	2013	2014	2015
🟩 Unstructured	11,430	16,737	25,127	39,237	59,600	92,536	147,885	226,716
🟨 Database	1,952	2,782	4,065	6,179	9,140	13,824	21,532	32,188
🟦 E-mail	1,652	2,552	4,025	6,575	10,411	16,796	27,817	44,091

It’s an undeniable trend: unstructured data is increasing exponentially, fueled by e-mail, eDocs, social media and Web content, as well as rich media content. The current industry term for this explosion is “Big Data.”

Now, more than ever, it is increasingly critical to deal with multi-source input before you find your organization buried beneath an avalanche of unstructured information.

BIG DATA OFFERS OPPORTUNITIES ALONG WITH THE CHALLENGES ❯❯❯

Big data isn’t just the volume of information, but the amount of data available for analysis, along with means of access and applicable technologies that can make greater sense of this data. You’ve got more input, and with smarter analytical tools, you can make more sense of it and turn this significant knowledge into better decisions.

Source: Enterprise Strategy Group, 2010.

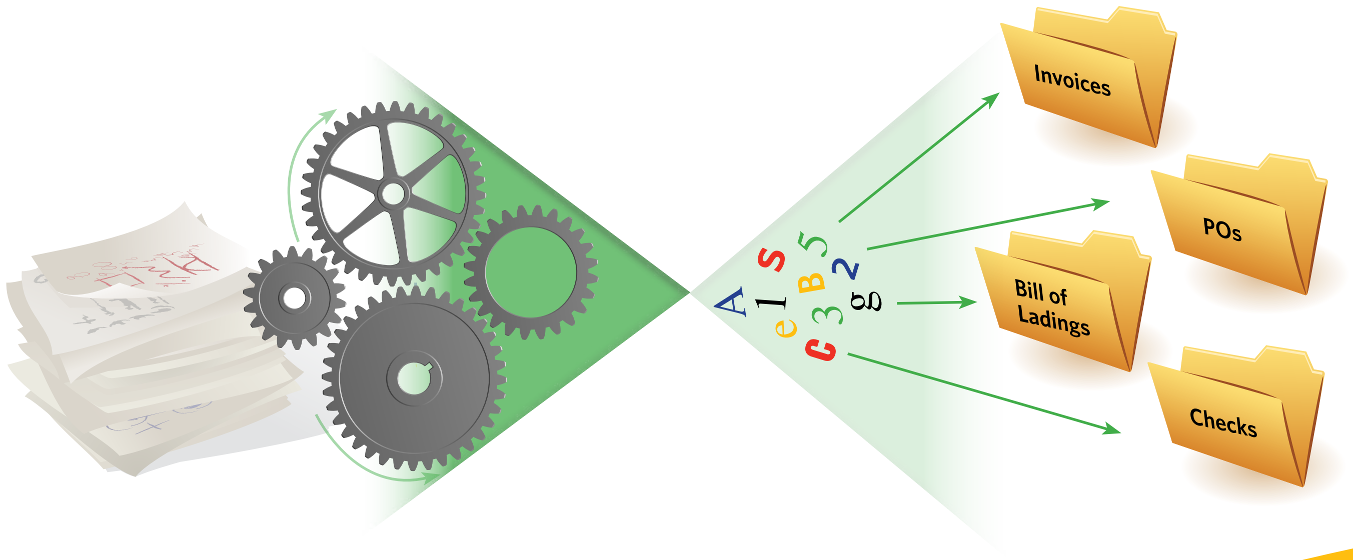
STEP 4 Document classification

The next step is the classification of a document. For paper, this replaces the sorting of incoming mail, where Larissa previously placed documents into bins for claims, receipts, new contract applications, and so on. In a computer system, these bins are folders.

Classification methods

There are roughly four methods of classifications, and these are sometimes combined:

- **Symbolic,** such as barcodes, is the easiest and most accurate. It typically applies to paper and fax sources, but not to e-mail and e-docs.
- **Graphics-based document analytics** work like a human visual categorizer, classifying documents by looking at their appearance but not reading text. For paper-based workflows, this is typically used for semi-structured type documents like invoices, where the document type is known.
- **Graphics- and text-based keyword combined.** Keyword is simple and typically combined with graphics-based analytics to add simple text rules. Example: when graphics says it is an invoice, and keyword says to look for the words "car repair," there is a high confidence about the classification. However, this process requires scripting and lacks flexibility.
- **Full, text-based document analytics** is the most recent and complex form of classification and the most versatile, as it classifies all text-based inputs - including OCR'd scanned paper, e-mails, e-docs and social media content.



KEY CONSIDERATION ❯❯❯

Ask your vendors about their approaches for document classification and if their solution can classify incoming e-mails – the short and long multi-threaded body – without keyword and complex rules. E-mail is one of the most challenging classification issues.

STEP 4

Document classification, cont.

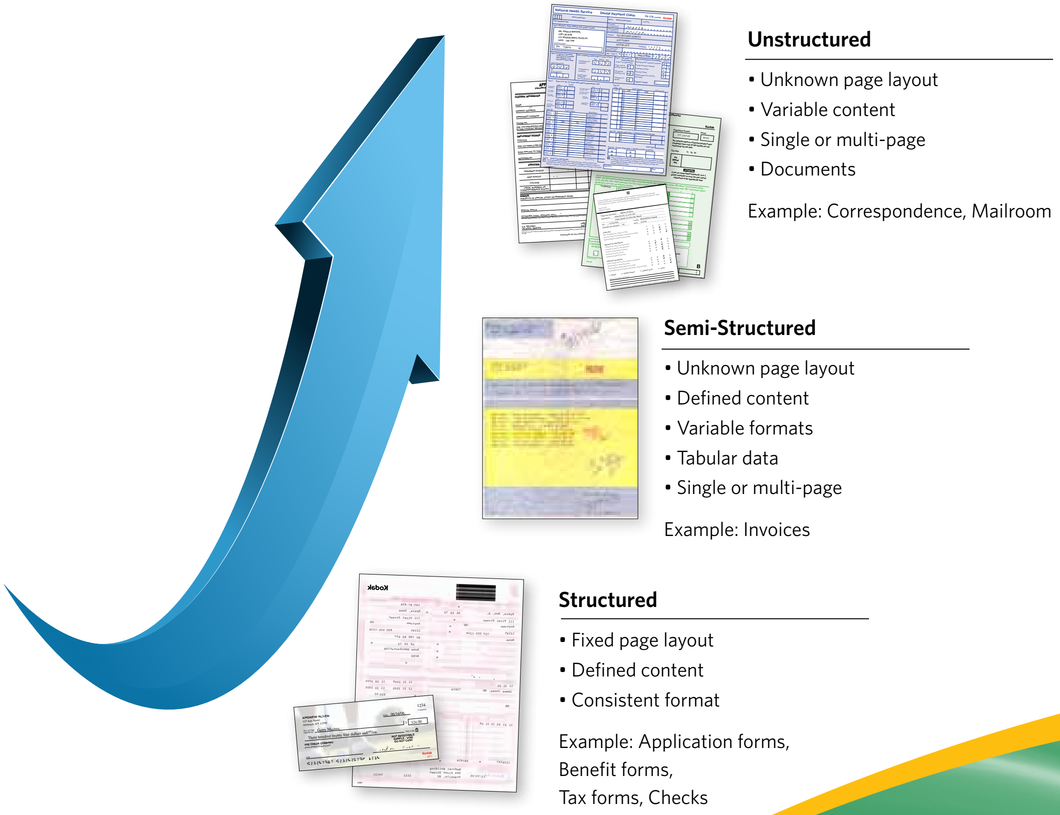
The importance of text-based classification techniques

A simplistic way of looking at the maturity of technology is to look at documents in three classes, structured forms, semi-structured documents and unstructured documents.

Start with structured forms, low in complexity with no variability. Forms are typically designed by the company, hence there is no need to classify the specific document type.

Next are semi-structured documents, such as invoices where the forms are of the same type but each vendor provides a different variation. Technology for the classification is typically graphics- and text-based, and processing of these documents, invoices in particular, is now also quite mature.

Unstructured document processing began a decade or so ago, yet remains a major challenge. Complexity and variability is extremely high and requires a new approach to classification. A small set of classes can be handled by rules and keywords but a more complex set requires text-based advanced classification techniques and algorithms that have been introduced as part of Machine Learning Science. These advances have arrived within the past five years, but most companies are still only applying classification based on scripted rules, like keywords.



Unstructured

- Unknown page layout
- Variable content
- Single or multi-page
- Documents

Example: Correspondence, Mailroom

Semi-Structured

- Unknown page layout
- Defined content
- Variable formats
- Tabular data
- Single or multi-page

Example: Invoices

Structured

- Fixed page layout
- Defined content
- Consistent format

Example: Application forms,
Benefit forms,
Tax forms, Checks

KEY CONSIDERATION ❯❯❯

Ask your vendors about the classification options for the different types of documents in your business.

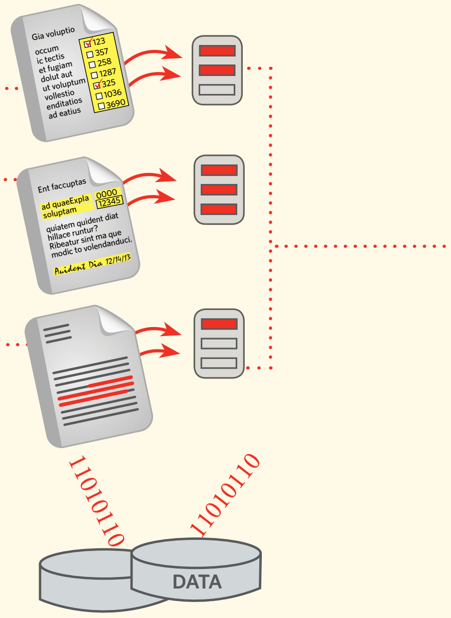
STEP 5 Data extraction

This is when information is derived via extraction of metadata from each document. This step replaces manual data entry and helps eliminate potential human errors. In our insurance example, knowledge worker Larissa receives the invoice from customer Jim for the repair of his car, but now needs to answer questions such as: what claim does this relate to? What is the amount? It's easy to see that this requires a time-consuming effort that slows down the handling of the claim significantly.

Data extraction methods

1. **Graphical**
   Based on visual observation – only viable for forms.
2. **Rules-based**
   Script rules include keyword anchors and regular expressions, such as "looking for claim number," "look for number next to claim," "x digits," and similar terms.
3. **Semantic understanding**
   Here areas of interest are highlighted and connected with a metadata field. This process is run on multiple documents (learning sets) so the system learns and can then act on the information in the future without human interaction.

Only a small portion of business situations can be adequately covered through scripted rules. The technology of semantic understanding is mature enough to be widely deployed, opening up huge opportunities for business process automation.



New frontiers in data extraction

Many vendors now tout **self-learning** – it's a new buzz word. However, what they often mean is that they make their rules a bit more adaptive, rather than using semantic understanding and machine learning.

Fuzzy understanding is part of best-of-breed solutions, because you cannot always take data literally. OCR makes mistakes. E-mails and documents can have misspellings. Few vendors offer fuzzy understanding, based on degrees of truth rather than absolutes.

Validation systems function similarly to a database lookup, validating a two-way match with the database system. Here, fuzzy understanding is still needed as the database may also contain errors, such as misspellings, duplicated entries with small variances, etc.

As an example: Larissa enters a claim number, insurance number, and customer name and address. With this information, the correct record can be found in the insurance system and validated.

If the data extraction system is not certain enough, the document needs to go to a validation station where Larissa will validate the data extracted. Some advanced systems also learn from validation input.

KEY CONSIDERATION ›››

Ask your solution provider about the approaches used for data extraction and how the specifics of self-learning techniques are applied.

STEP 6

The Eight Steps of Intelligent Document Understanding

Enable document collaboration

"Search-ability" and "share-ability" were noted as key reasons for document digitization by survey respondents in Step 1. These same aspects are critical for multi-source documents. As Gartner Research states, **"People don't want computers. They want to relate, share, communicate, enjoy, learn, discover, analyze and create.[2]"** And they want to optimize the power and value of their documents and knowledge wherever and whenever – on any device ranging from laptop to tablet to smartphone, and new hybrids to come.

A Content Management System provides a platform for users to search and collaborate with documents. Think **Microsoft SharePoint, Alfresco** and **Box.** Microsoft President Steve Ballmer defined **SharePoint** as "a general-purposed platform for connecting people with information." And there are others, such as **Microsoft** Yammer, **IBM** Connections, and **Salesforce Chatter.**

AUTOMOBILE CLAIM REPORT
Agent copy

Reporting agent: Bryce Bilton - Agent   Agent code: 54-567   Agent phone: (555) 456-7890
Claim: Claim office: Benton
Claim rep: Claim rep phone:

INSURED
Coverage: M54J, C12, F480 GLASS, 1245, C 55/222, 10B
Policy number: 845 5456C56-74M   Insured: Bonnie Boyd
Date of loss: 09-24-05   Address: 545 Avenue B
Time of Loss: 5:00 PM   City: Benton   St: FL   Zip: 54567
Date reported: 09-24-05   Phone: 555-545-6789
Contact: Bonnie
Location of Loss: Bugsy Street   City: Benton   St: FL

VEHICLE 1
Insured vehicle/year/make/model/bodystyle: 00 NISSAN ALTIMA 4 DR
VIN: JUDOG8456EX224400   License number/state: BCD890/FL   Involved in loss? YES
Prior damage:
Principal damage: PASSENGER DOORS, FRONT AND BACK
Driver: Boyd, Bonnie
Age: 55   Date of birth: 10-01-1975   Sex: F
Injuries: None

VEHICLE 2
Year/make/model/bodystyle: 98 MAZDA 626 4 DR
License number/state: DEF910 /FL
Principal damage: FRONT END BUMPER
Location:
Insurance co: Bee's Insurance Co.   Policy number: 2567890-11
Owner/driver: Bither, Brent   Address: 545 B Street
City: Benton   St: FL   Zip: 54567
Phone: 890-2545

FACTS
VEH 1 WHICH IS OUR INSURED'S VEH WAS STRUCK FROM THE SIDE BY VEH 2.   Police report made? YES
Dept where reported: Benton
Report number: VE-2545

Insured violation?   Describe:
Claimant violation?   Describe:

Policy number: 845 5456C56-74M   Insured's Initials □
Date of Loss: 09-24-05

"Share-ability" across devices and platforms.

Collaboration in the real world

Revisiting the insurance scenario: customer Jim may be able to see his case on a portal, then upload a video to support his claim via the Content Management System. Dave, the insurance agent, could come to Jim's house and review and edit the case document set on his iPad at Jim's kitchen table. Now everyone has access to the same up-to-date content and information. This is practical, powerful, and offers significant time- and dollar-saving benefits.

KEY CONSIDERATION ❯❯❯

Analyze your processes and identify opportunities to enhance collaboration for critical business processes.

[2] Reference: Gartner Research, 2012.

STEP 7

The Eight Steps of Intelligent Document Understanding

Automatically initiate business transactions

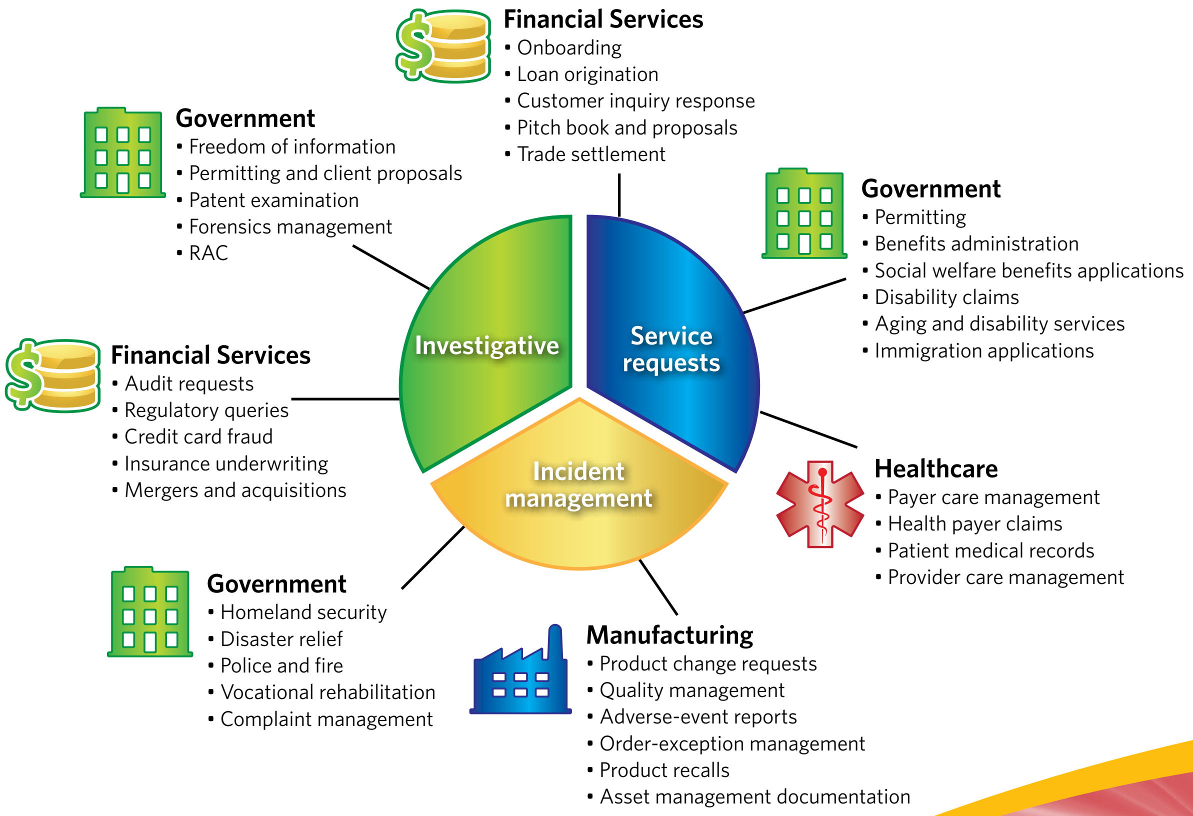
Documents with meaningful metadata – residing in a powerful Content Management System – allow automatic triggering of business workflows, and make information actionable.

In our insurance example, once classified, the document will be automatically assigned to the claim case and the workflow can check for completeness of the document set, initiating communication with customer Jim to request a missing form. Similarly, an insurance claim approval workflow can be automated to interact with the insurance system and enable appropriate payment. Just imagine how this will greatly reduce time involved in handling the claim!

And that's just the beginning

Combine an intelligent document and a modern ECM solution with workflow capability, and the applications and advantages multiply.

Opportunities for Application Automation



Financial Services
- Onboarding
- Loan origination
- Customer inquiry response
- Pitch book and proposals
- Trade settlement

Government
- Freedom of information
- Permitting and client proposals
- Patent examination
- Forensics management
- RAC

Government
- Permitting
- Benefits administration
- Social welfare benefits applications
- Disability claims
- Aging and disability services
- Immigration applications

Financial Services
- Audit requests
- Regulatory queries
- Credit card fraud
- Insurance underwriting
- Mergers and acquisitions

Healthcare
- Payer care management
- Health payer claims
- Patient medical records
- Provider care management

Government
- Homeland security
- Disaster relief
- Police and fire
- Vocational rehabilitation
- Complaint management

Manufacturing
- Product change requests
- Quality management
- Adverse-event reports
- Order-exception management
- Product recalls
- Asset management documentation

Source: Forrester, Dynamic Case Management; Definitely Not Your Dad's Old School Workflow/Image System, 2011.

KEY CONSIDERATION ❯❯❯

Ask your solution provider about the options to automatically start a process. While the business advantage is significant, it is also key to have good control mechanisms in place to ensure the proper execution of the transaction, and the involvement of the knowledge worker when necessary.

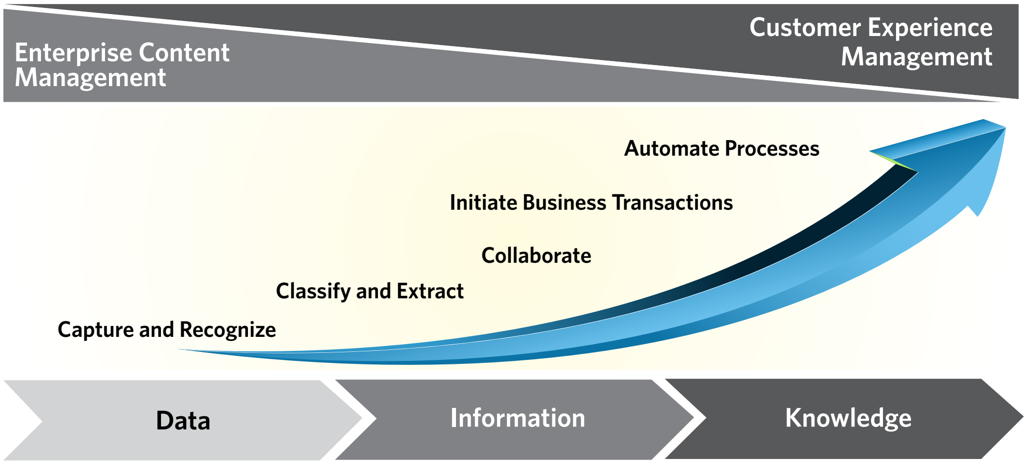
STEP 8 Automate business processes with advanced insight

With an intelligent document capture system feeding content and metadata into a Content Management System, tasks that previously required manual intervention are now automated for knowledge worker Larissa, thanks to scripted rules-based workflow.

So what happens when the multi-source system receives an e-mail from customer Jim to Larissa with an address change? It's actually a difficult problem to detect an address change in an e-mail, as there is no validation record and the e-mail may contain both old and new addresses. Thanks to intelligent data extraction, the address change triggers a scripted workflow to correct the address field in the insurance company's database.

To add another level of advantages, enable the Content Management System with a Knowledge Management System featuring analytics capabilities for big data. Now the Knowledge Management System explores multiple content systems – including social media – and discovers competitive information about Jim's new community showing that a competitor is very strong in that region. This creates an alert that's sent to Chuck, the insurance agent near Jim's new address, who calls Jim to make sure he is happy with his coverage, and to proactively preempt a competitor's sales approach.

This is state-of-the-art advanced intelligence at work, as all of these conditions and variables could not be scripted. Today, using Business Intelligence on structured data (a database) is becoming current, but using it on unstructured data *is the future* and can only occur via intelligent document understanding.



KEY CONSIDERATION ❯❯❯

Evaluate your processes not only with the Enterprise Content Management aspect in mind, but consider the end-to-end process including aspects of customer experience management. This will allow you to identify the entire opportunity for the enhancement of your processes. Ideally, investments in future solutions should provide capabilities in this area.

Moving from knowledge to wisdom

Wisdom lets you turn knowledge into predictive analytics to help drive your business, create new opportunities, and streamline processes. Larissa now benefits from all these advantages – spending less time doing data entry and case management. She's more involved with customers and better understands their evolving needs. Now she's become a proactive knowledge worker, with the time to focus on sustaining and building business and relationships. And her insurance company employer is delighted because meaningful time and money savings are being achieved, relationships are being enhanced through greater customer satisfaction, and multi-source data is far more aggregated, meaningful, and accessible.

Kodak's Document Imaging business enables customers to capture and manage valuable information from electronic and paper documents. Our solutions include award-winning scanners and capture software, information workflow software, an expanding range of professional services and industry-leading service and support. From small offices to global operations, Kodak has the solutions to automate your business processes and intelligently deliver the information your enterprise needs.

To learn more:
www.kodak.com/go/docimaging

Printed using **Kodak** Technologies.

Eastman Kodak Company
343 State Street, Rochester, NY 14650   1-800-944-6171

Kodak Canada, Inc.
Toronto, Ontario M9R 0A1   1-800-465-6325

And, because you never know
where technology may take us,
Bob may become super Bob if he can learn
to predict the outcome of the next World Cup!