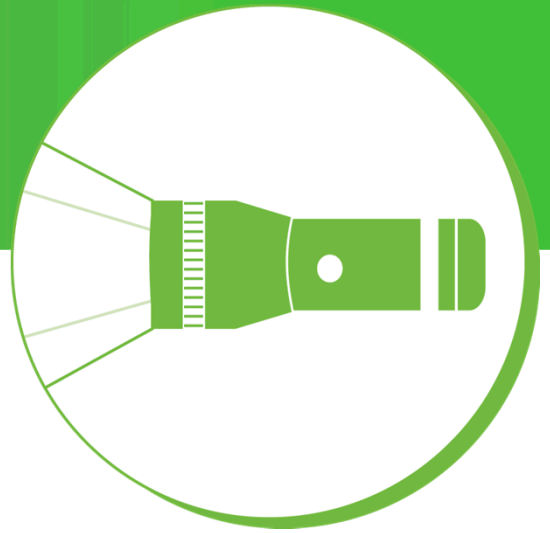
# ECM and Dark Data

*Turn on the light to improve compliance, cut storage, and leverage document assets*

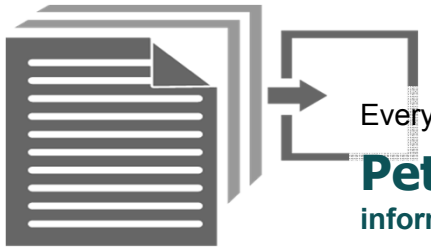
Peter Duff
CEO, Adlib
Vice Chair, PDF Association

Roger Beharry Lall, ecm$^{P}$
Director, Product Marketing, Adlib



# The Data Explosion

## INDIVIDUALS ARE CREATING VAST AMOUNTS OF DATA.



Every day, **15 Petabytes** of **new information** is created


By 2020, **B2B transactions** on the internet will reach **450 billion per day**

## WHAT'S CONTRIBUTING TO THE EXPLOSION?

By 2015, nearly **3 billion** **people** will be online, **creating and sharing 8 zettabytes**

**File Type Growth Rate of Consumer Internet Traffic**

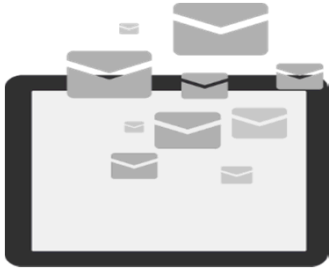
File Sharing **23%**

Data **29%**

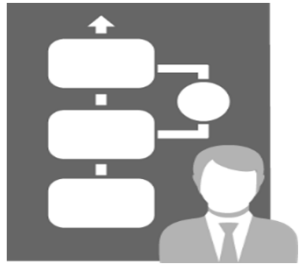
(CAGR 2010-2015)

**Enterprise data will grow** **650%**, partially due to regulations like the Sarbanes-Oxley Act requiring companies to store financial records

## DATA... WE HAVE A PROBLEM.



**35%** of the digital universe is **subject to compliance and regulations**

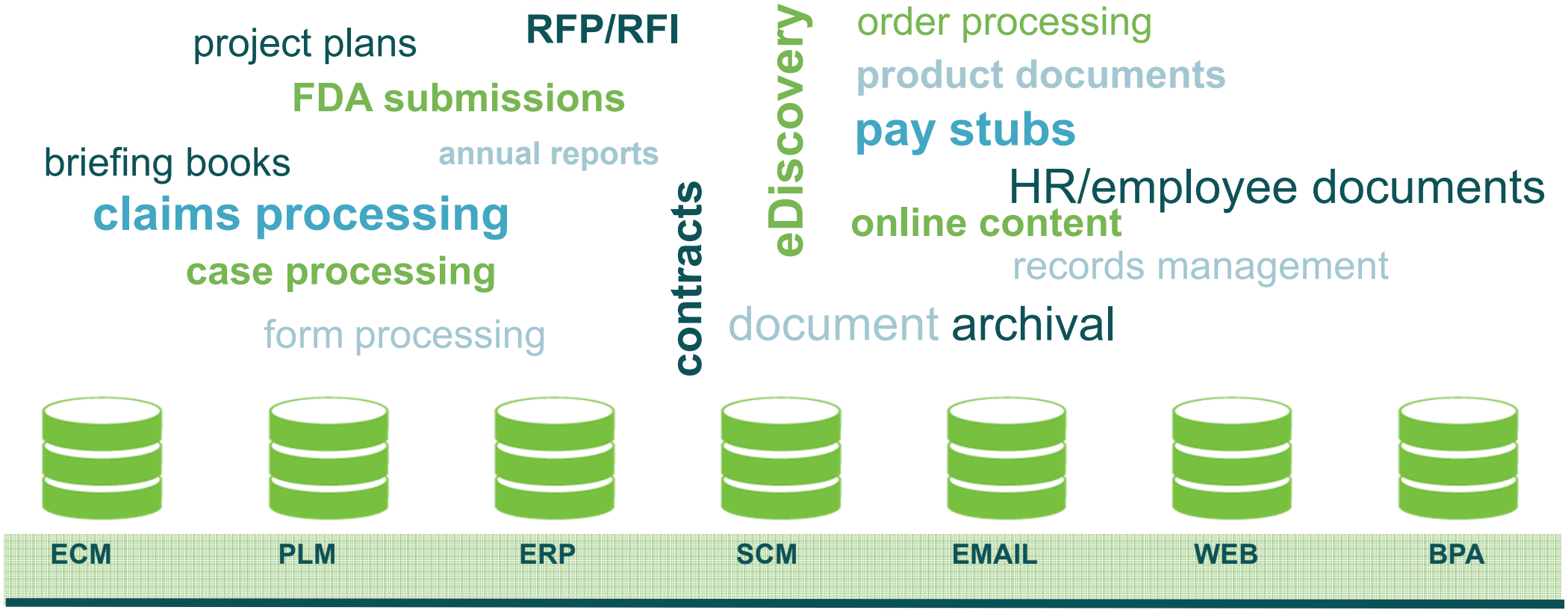


In the next decade, the number of files will grow by a factor of **75** while **IT professions** will grow by **less** than a factor of **1.5**

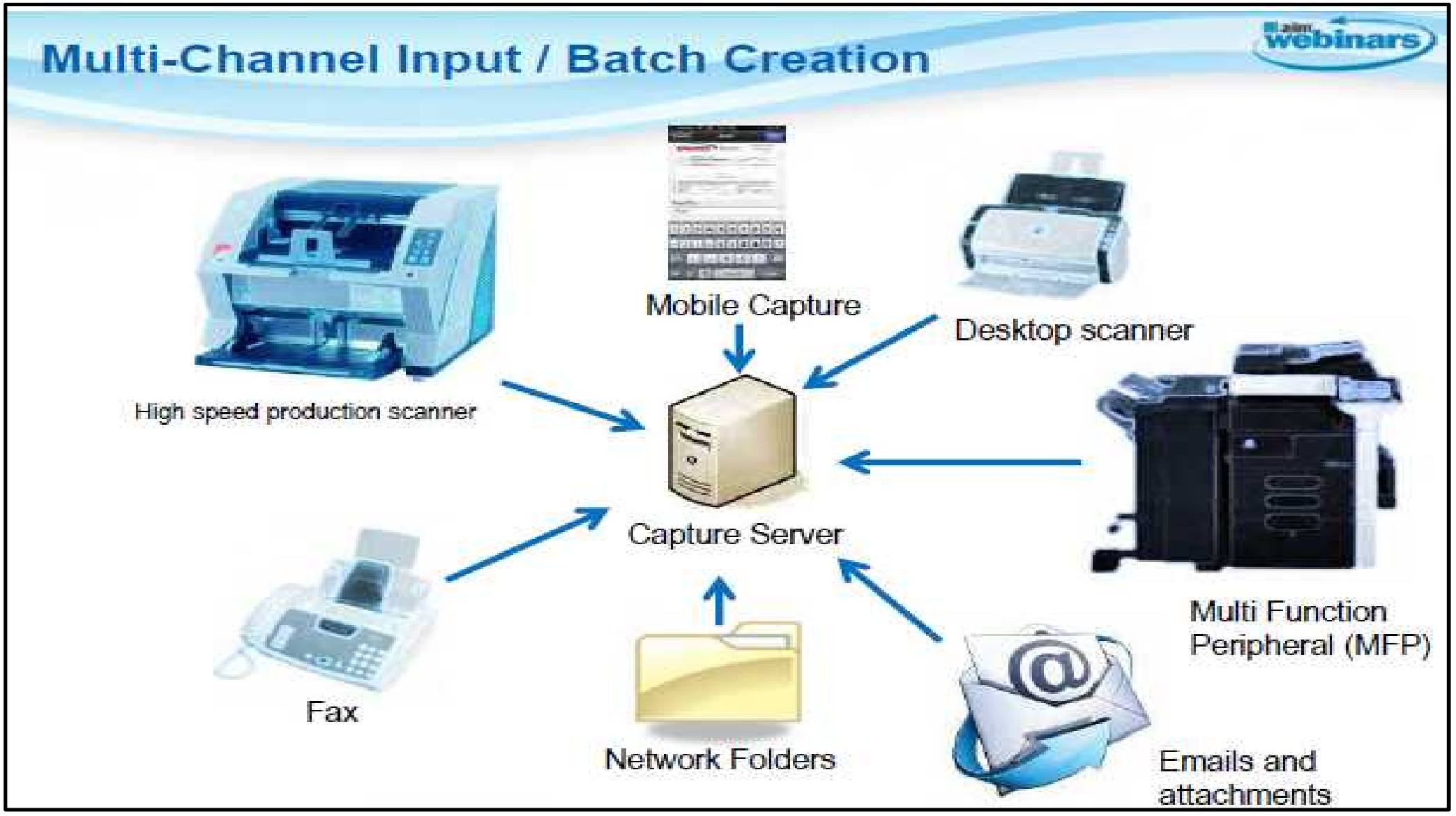
ADLIB®

# Document Complexity

## Variety Of Systems, Processes And Formats



# Multi-Channel Input / Batch Creation




Source: AIIM – 2014 via Harvey Spencer

# Growth of Dark Data


ADLIB®

*Source: HP Syncsort*

Gartner

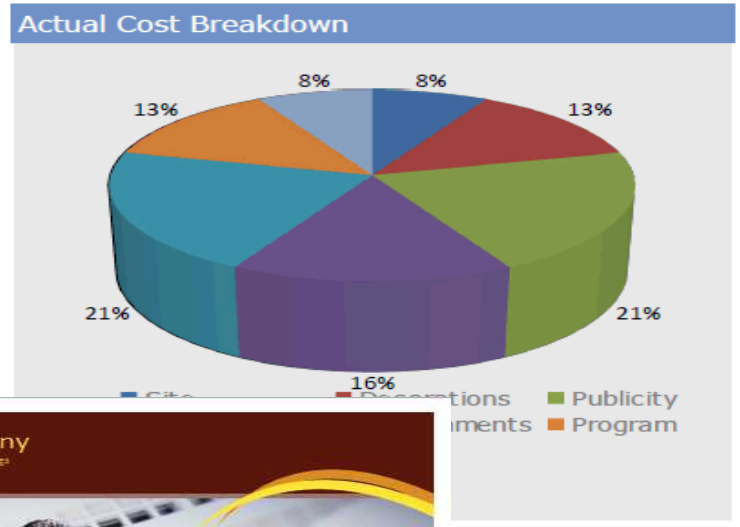
**The information assets that organizations collect, process and store during regular business activities, but generally fail to use for other purposes.**

IDC

**Up to 90% of Big Data is Dark Data.**

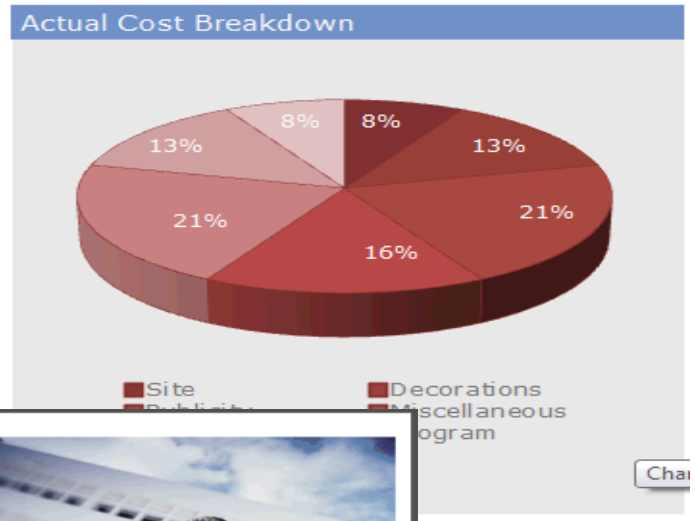
ADLIB

# Dirty Data. Dark Data. No Data!





Original Source




Rendered Image

ADLIB


RISK
RISKS
MANAGEMENT
EDIT
MAY
PROJECT
EVENT
COST
VALUE
ENTERPRISE
ALSO
FINANCIAL
ANALYSIS
MITIGATION
PRACTICE
MEASURES
CASE
EVENTS
TRANSFER
COMPANY
OCCUR
ONE
OPERATIONAL
PROCESSES
PROCESS
THREATS
OFTEN
POTENTIAL
AVAILABLE
MANAGING
LOSS
PROBABILITY
INFORMATION
IMPLEMENTATION
ANOTHER
HAZARD
ASSESSED
FIRST
REDUCE
ISO
RESOURCES
IDENTIFICATION
EXAMPLE
TIME
CONSEQUENCES
METHODS
POSSIBLE
ACTIVITIES
METHOD
ORGANIZATION
TECHNOLOGY
SOFTWARE
USED
IDENTIFIED
SOURCE
PLAN
IMPACT
BUSINESS
DEVELOPMENT
COMMUNICATION
FOLLOWING
OCCURRENCE
INSURANCE
CONTROLS
ASSESSMENT
REDUCTION
HIGH

# Enterprise Content Management: Only Part of the Answer

## Enterprise

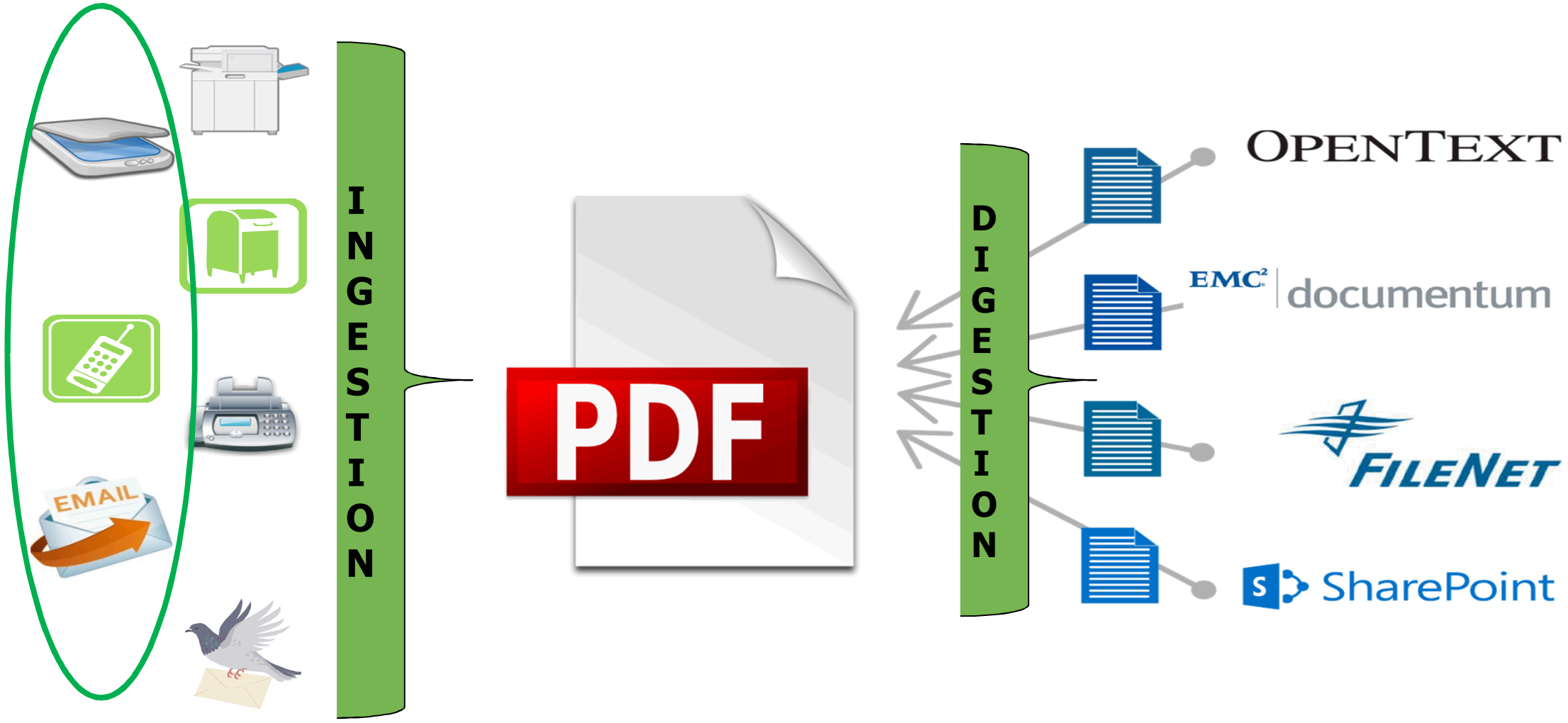
- Scalability
- High Availability
- Fault Tolerance
- Cloud/Virtualization

**Content**
**??**


## Management

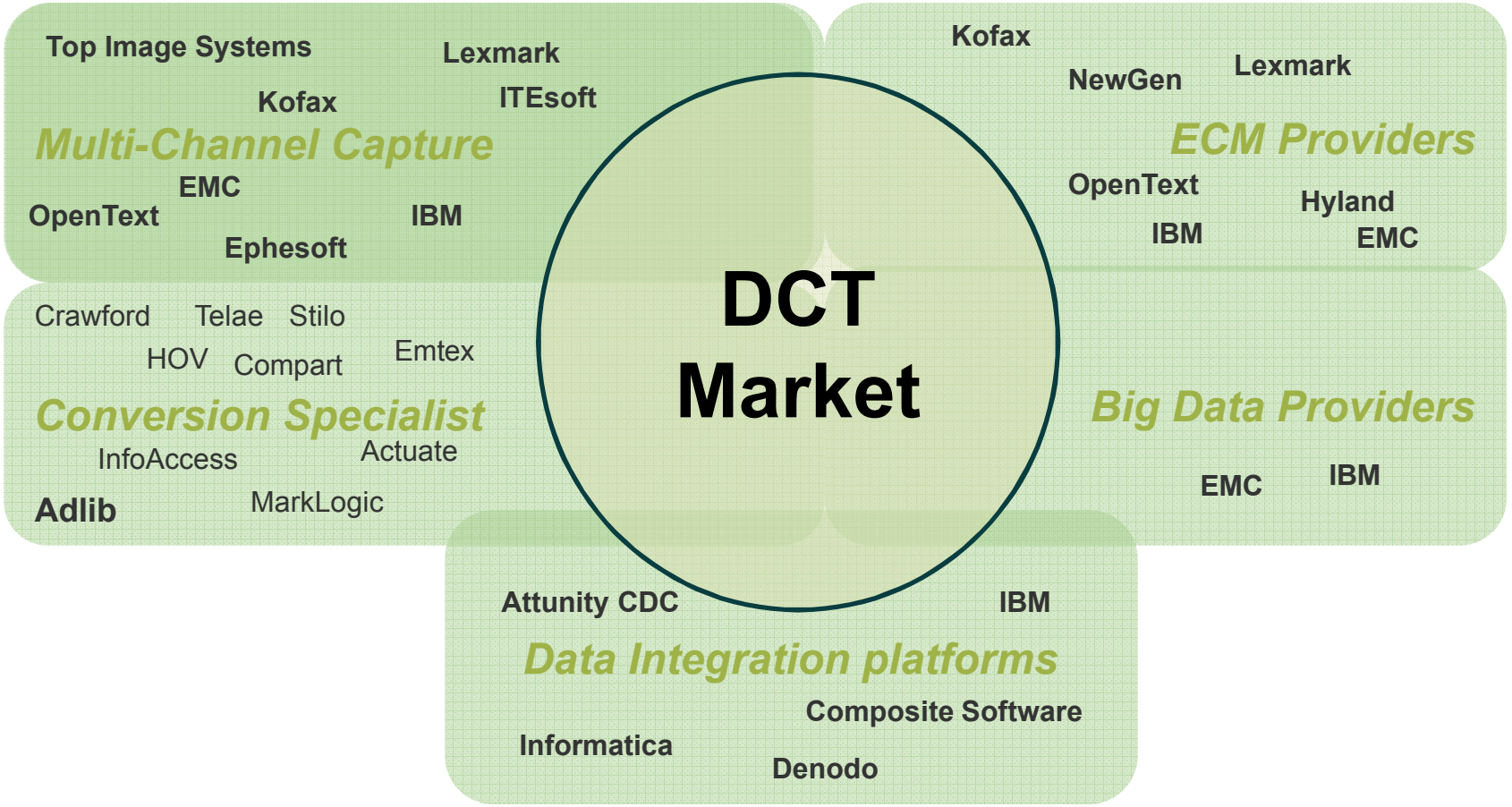
- Taxonomy
- Rules
- Permissions
- Metadata

ADLIB®

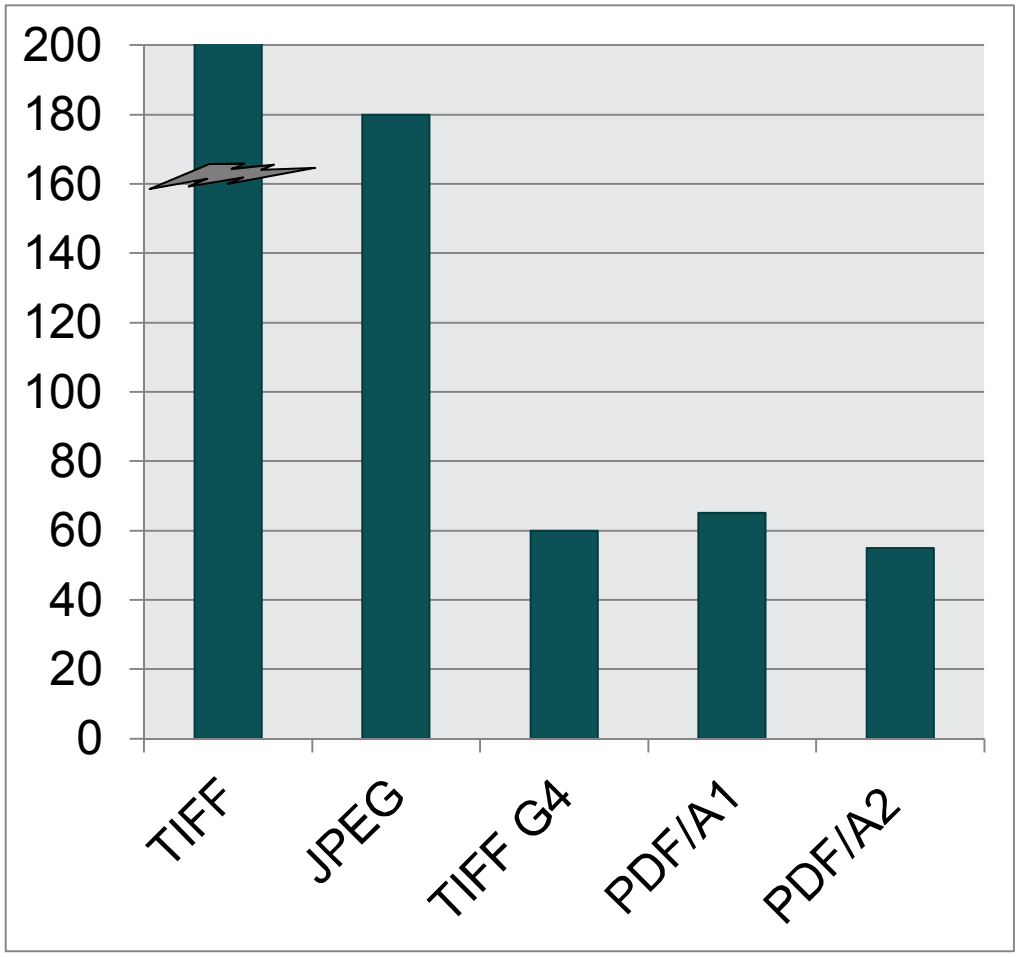

EMAIL
INGESTION
PDF
DIGESTION
OPENTEXT
EMC² documentum
FILENET
SharePoint

ADLIB®

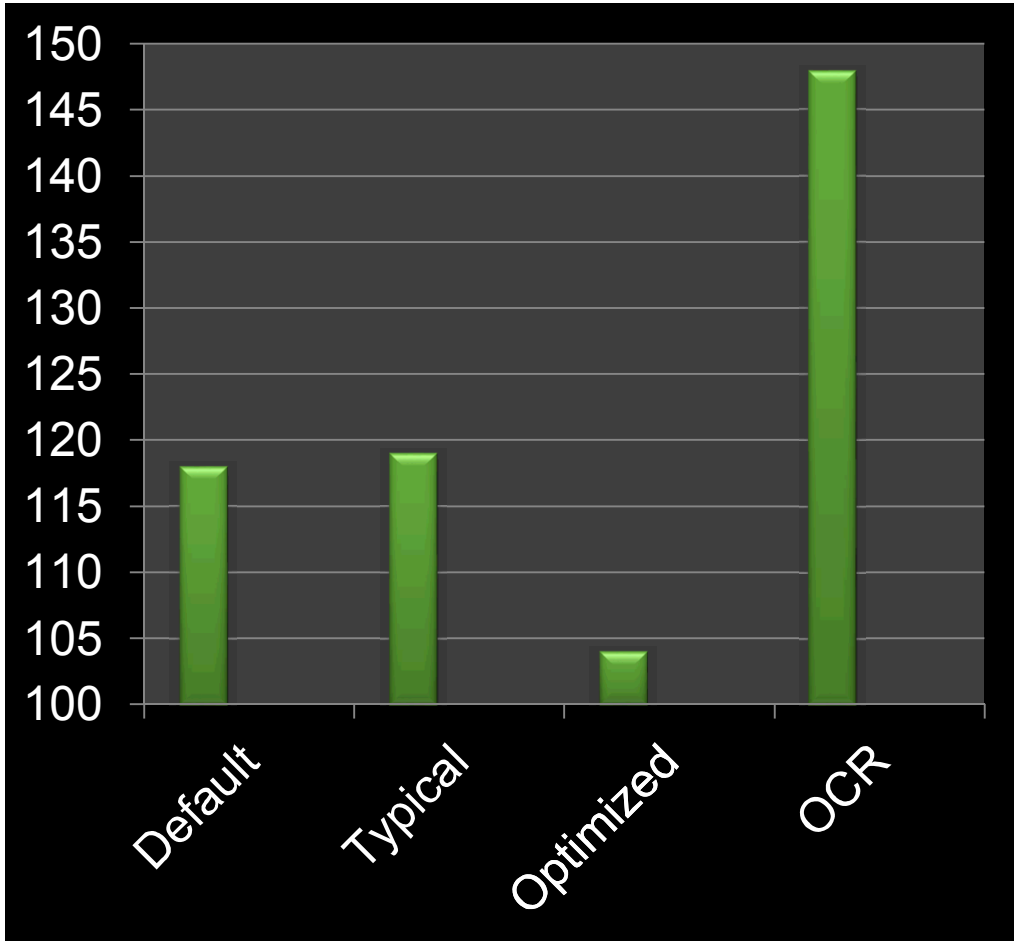
# Document Content Transformation



# File Size Optimization for Storage Reduction



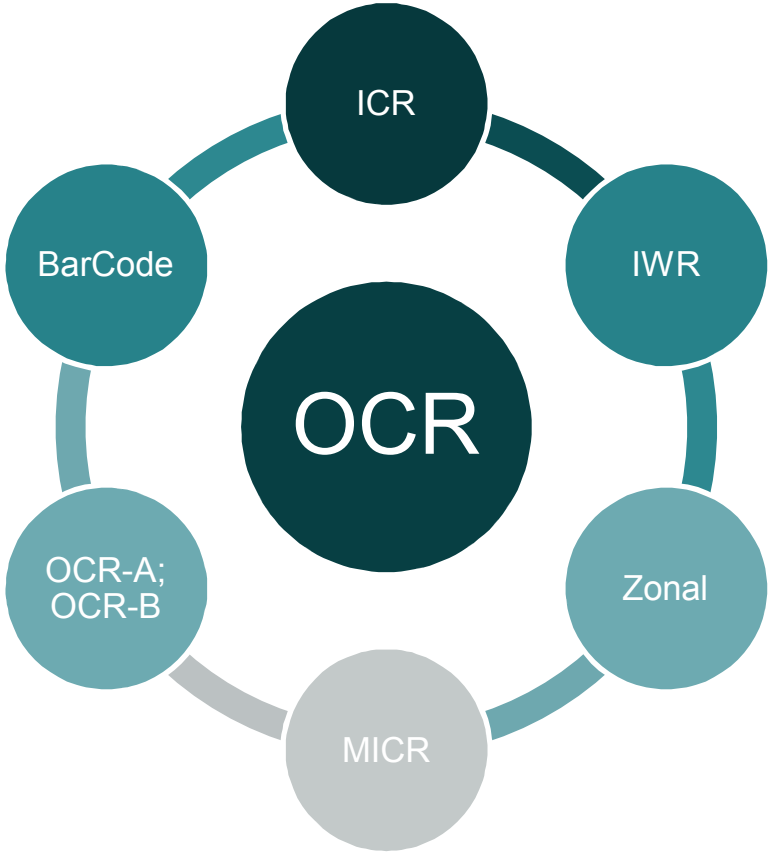

OPTIMAL FORMAT




OPTIMAL PDF

ADLIB

# Optical Character Recognition (OCR)



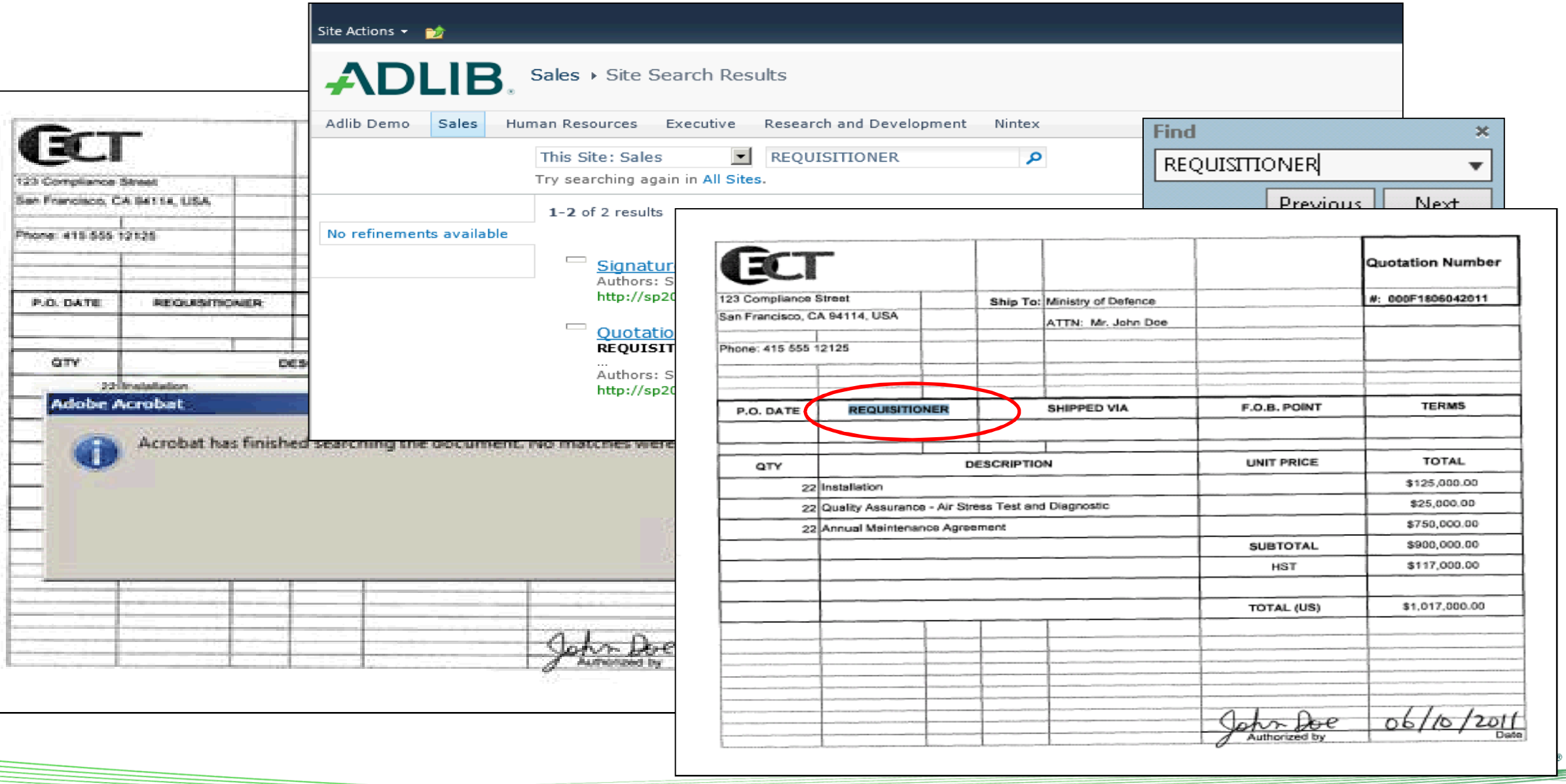

Converting printed or written text characters—captured as images during scanning—into computer-based, encoded text.

Benefits of OCR Capabilities

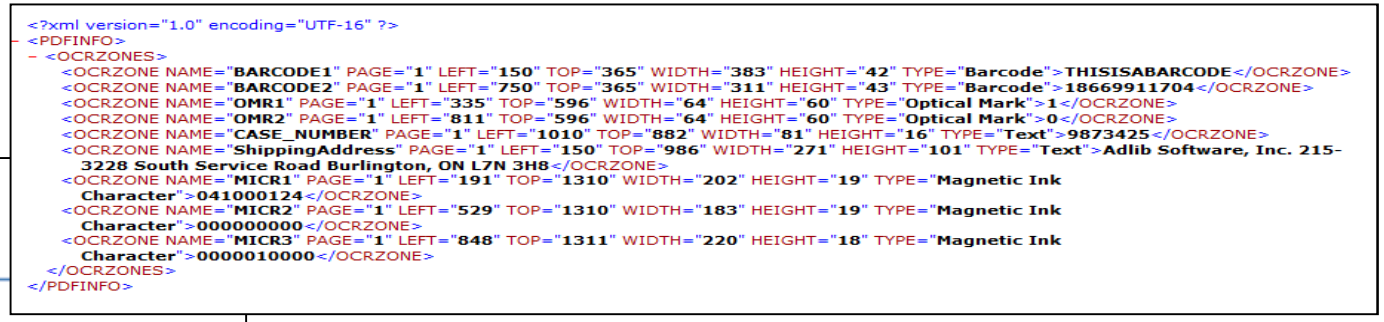
- Liberating information for electronic searches
- Delivering industry-leading accuracy
- Supporting regulatory mandates
- **Make content immediately findable from the moment of capture.**

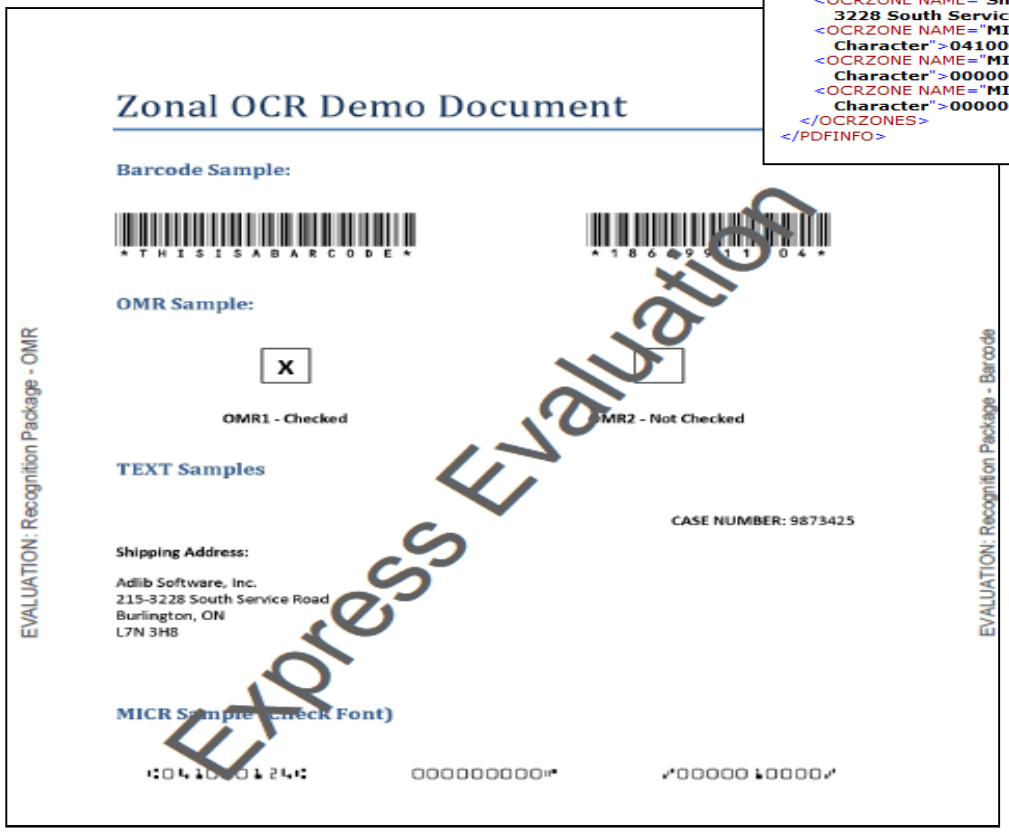
ADLIB®

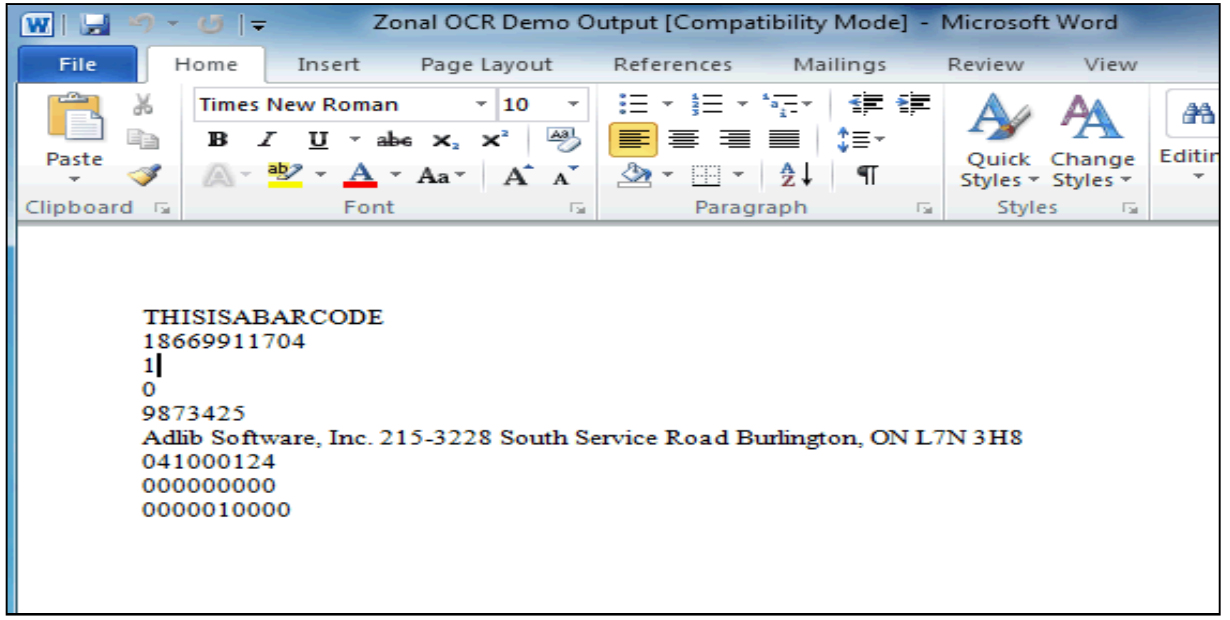
# Search Enabled Documents



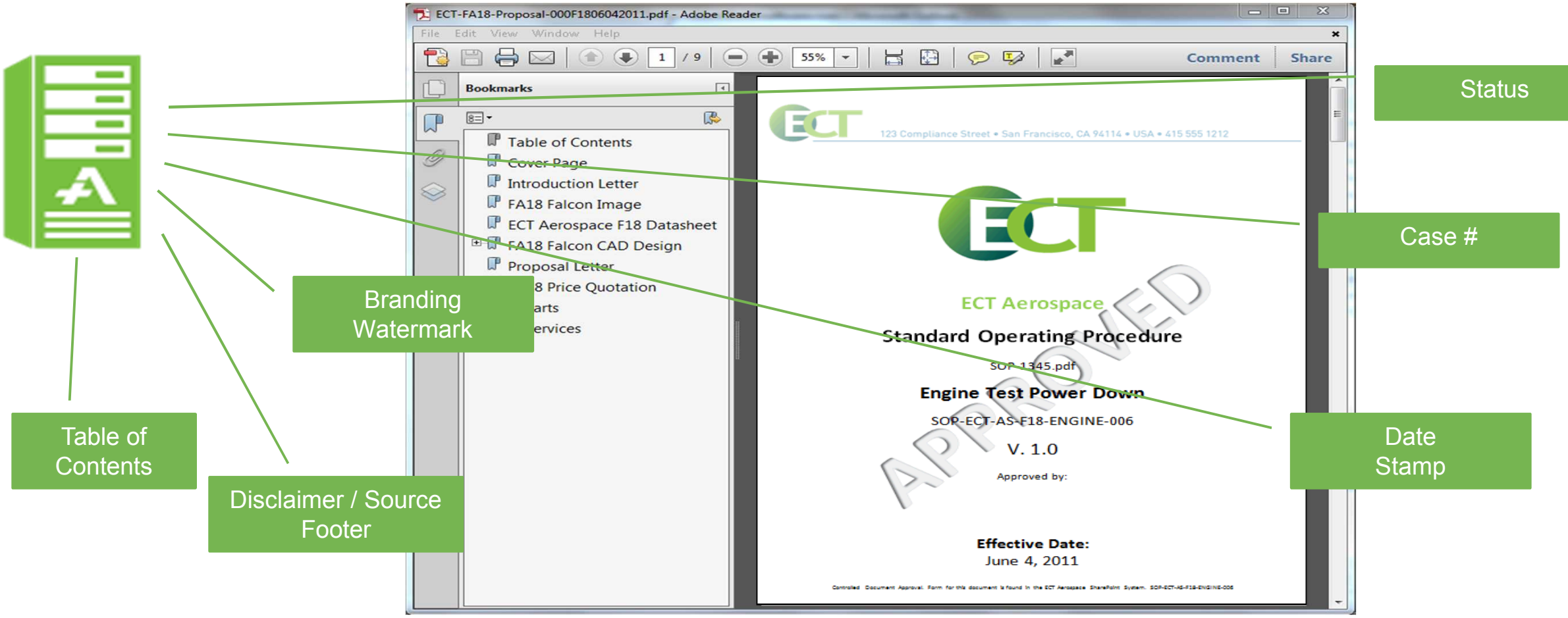
# Zonal OCR


<?xml version="1.0" encoding="UTF-16" ?>
<PDFINFO>
<OCRZONES>
<OCRZONE NAME="BARCODE1" PAGE="1" LEFT="150" TOP="365" WIDTH="383" HEIGHT="42" TYPE="Barcode">THISISABARCODE</OCRZONE>
<OCRZONE NAME="BARCODE2" PAGE="1" LEFT="750" TOP="365" WIDTH="311" HEIGHT="43" TYPE="Barcode">18669911704</OCRZONE>
<OCRZONE NAME="OMR1" PAGE="1" LEFT="335" TOP="596" WIDTH="64" HEIGHT="60" TYPE="Optical Mark">1</OCRZONE>
<OCRZONE NAME="OMR2" PAGE="1" LEFT="811" TOP="596" WIDTH="64" HEIGHT="60" TYPE="Optical Mark">0</OCRZONE>
<OCRZONE NAME="CASE_NUMBER" PAGE="1" LEFT="1010" TOP="882" WIDTH="81" HEIGHT="16" TYPE="Text">9873425</OCRZONE>
<OCRZONE NAME="ShippingAddress" PAGE="1" LEFT="150" TOP="986" WIDTH="271" HEIGHT="101" TYPE="Text">Adlib Software, Inc. 215-3228 South Service Road Burlington, ON L7N 3H8</OCRZONE>
<OCRZONE NAME="MICR1" PAGE="1" LEFT="191" TOP="1310" WIDTH="202" HEIGHT="19" TYPE="Magnetic Ink Character">041000124</OCRZONE>
<OCRZONE NAME="MICR2" PAGE="1" LEFT="529" TOP="1310" WIDTH="183" HEIGHT="19" TYPE="Magnetic Ink Character">000000000</OCRZONE>
<OCRZONE NAME="MICR3" PAGE="1" LEFT="848" TOP="1311" WIDTH="220" HEIGHT="18" TYPE="Magnetic Ink Character">0000010000</OCRZONE>
</OCRZONES>
</PDFINFO>

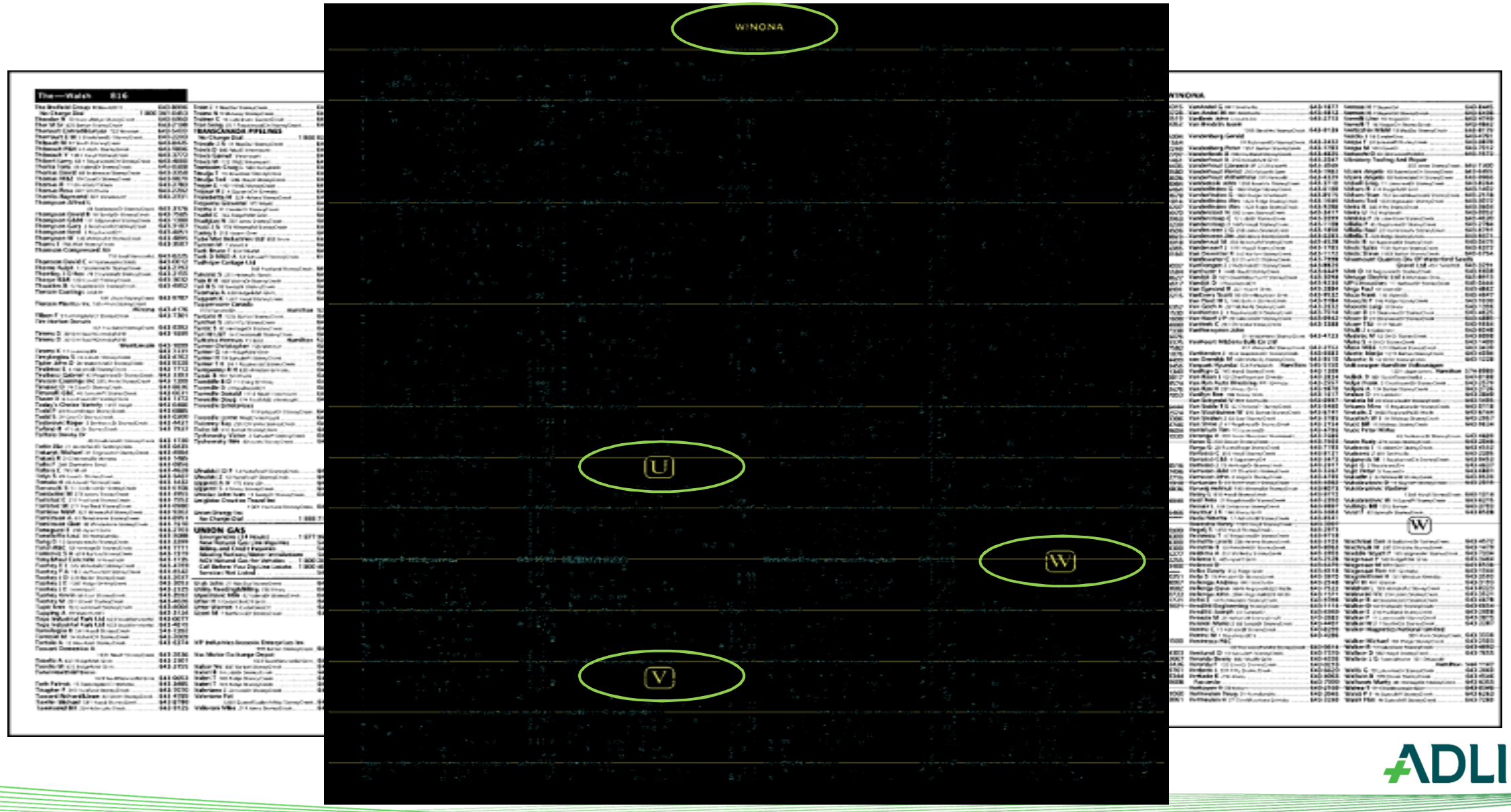

Zonal OCR Demo Document
Barcode Sample:
*THISISABARCODE*
*18669911704*
OMR Sample:
X
OMR1 - Checked
OMR2 - Not Checked
TEXT Samples
CASE NUMBER: 9873425
Shipping Address:
Adlib Software, Inc.
215-3228 South Service Road
Burlington, ON
L7N 3H8
MICR Sample (Check Font)
EVALUATION: Recognition Package - OMR
EVALUATION: Recognition Package - Barcode
Express Evaluation


Zonal OCR Demo Output [Compatibility Mode] - Microsoft Word
File
Home
Insert
Page Layout
References
Mailings
Review
View
Times New Roman
10
Paste
Clipboard
Font
Paragraph
Quick Styles
Change Styles
Styles
THISISABARCODE
18669911704
1
0
9873425
Adlib Software, Inc. 215-3228 South Service Road Burlington, ON L7N 3H8
041000124
000000000
0000010000

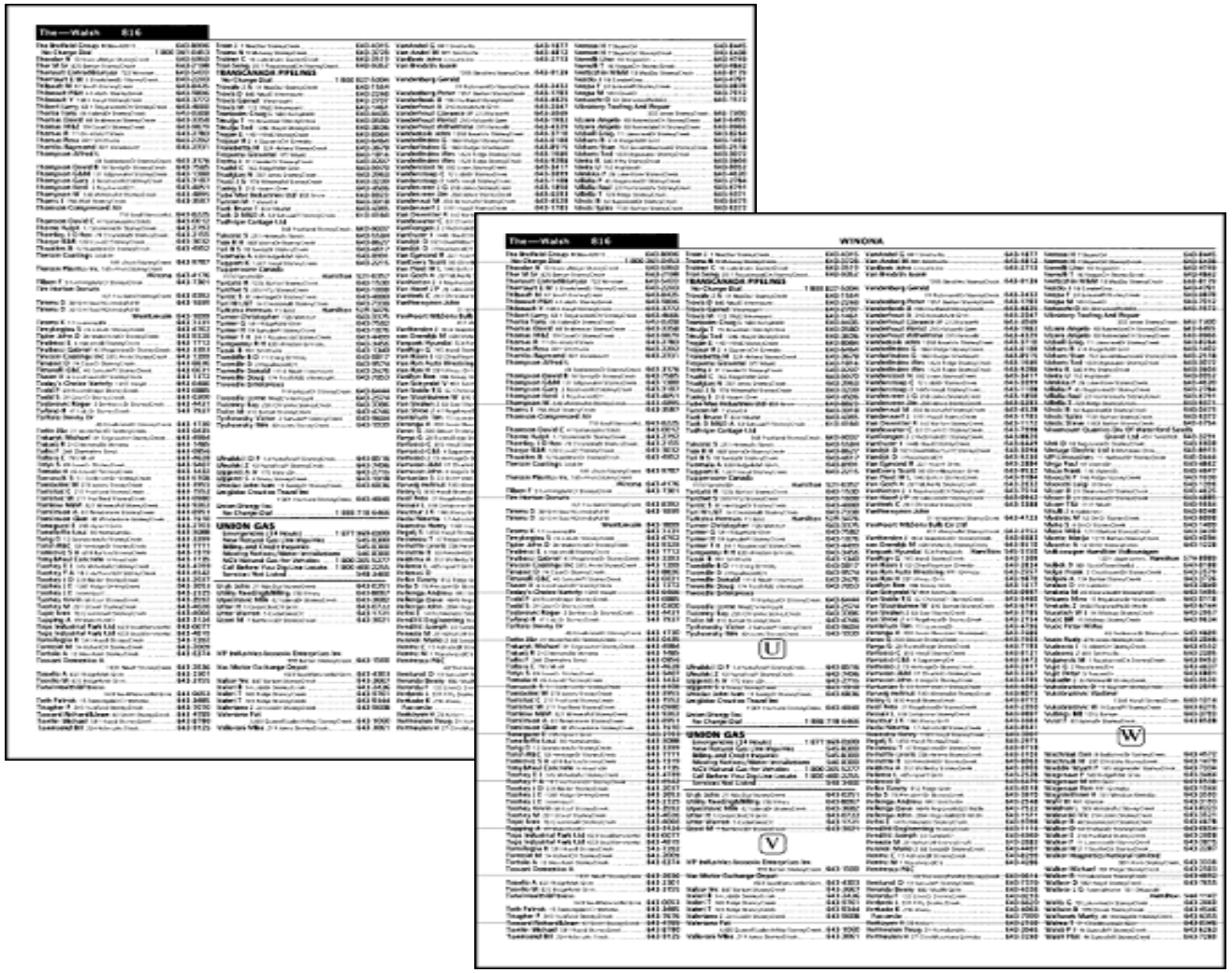
# Metadata Extraction



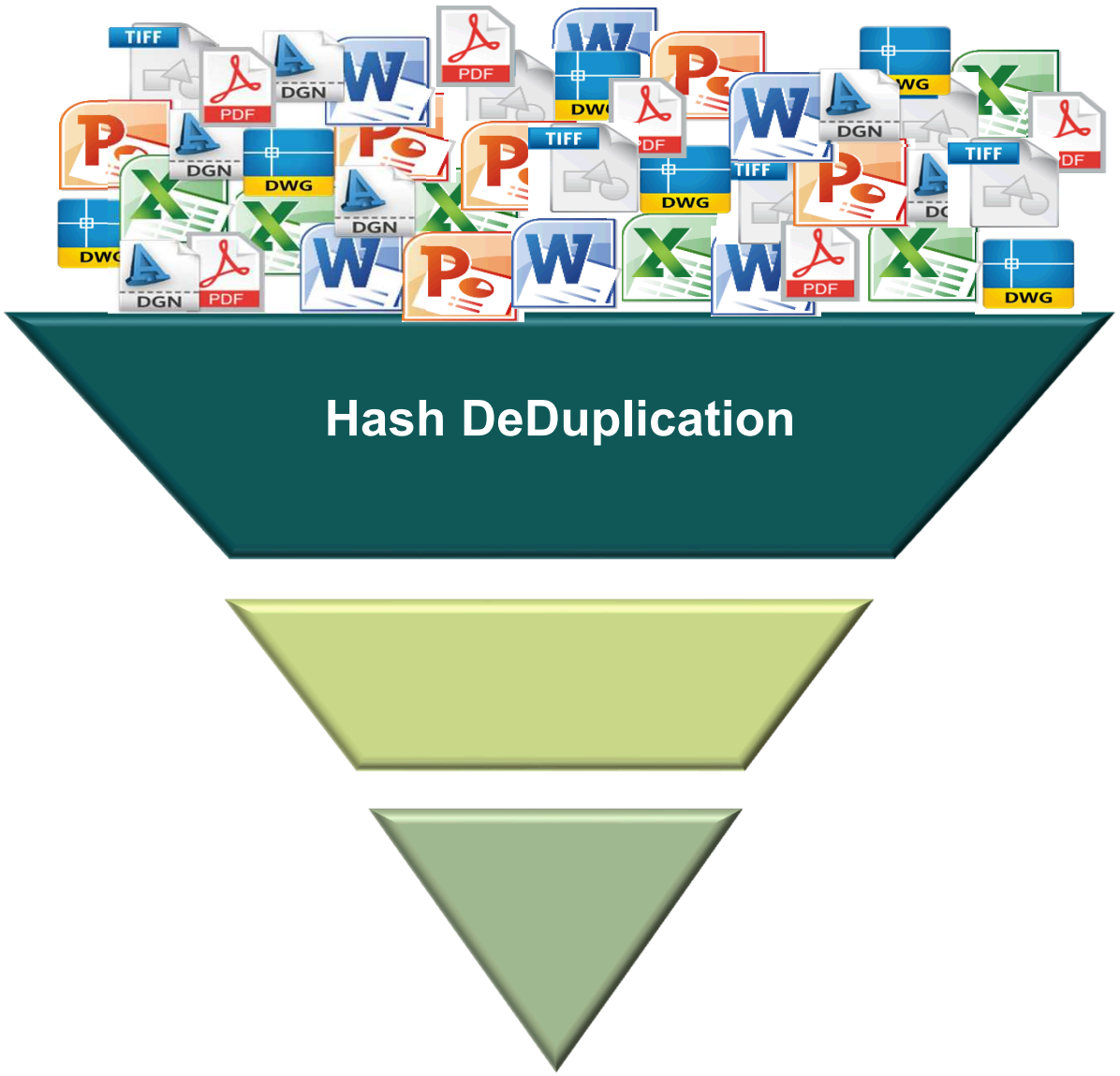
# Content Comparisson



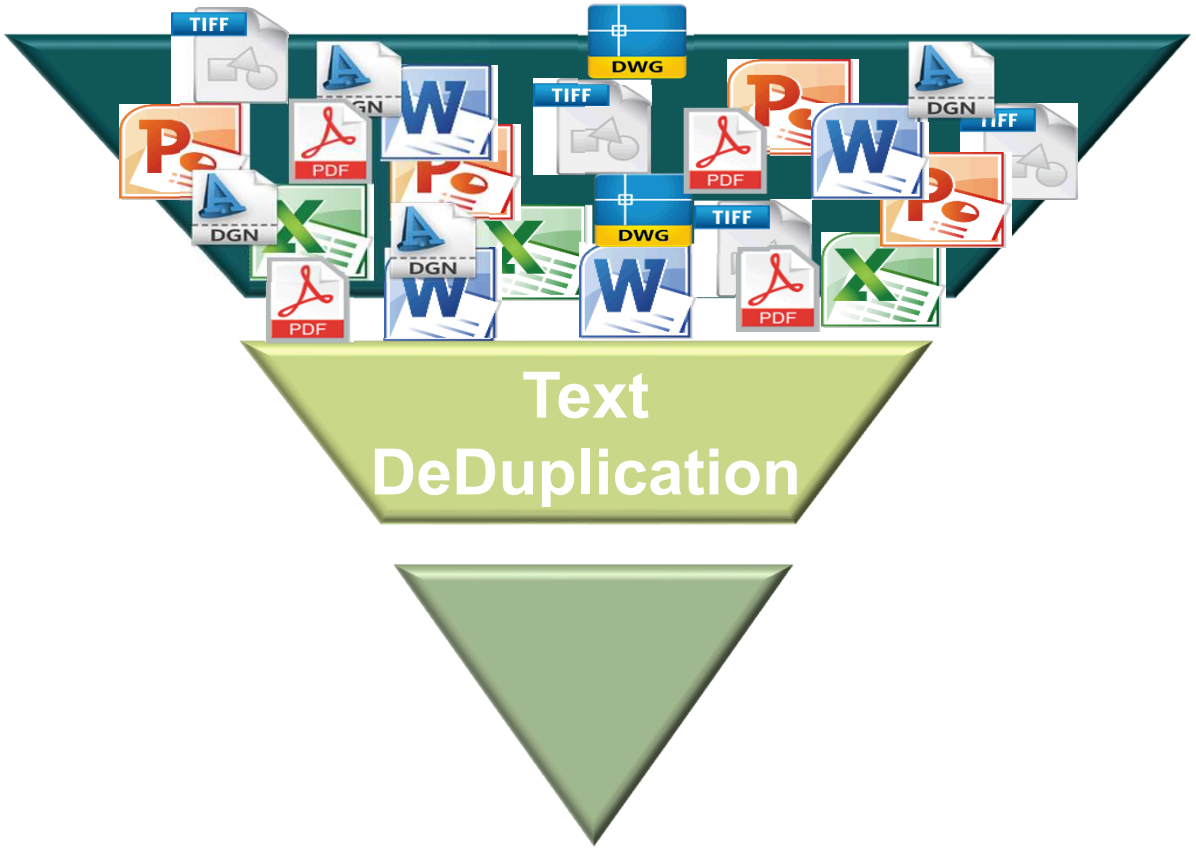
# Applications of Image Analysis:



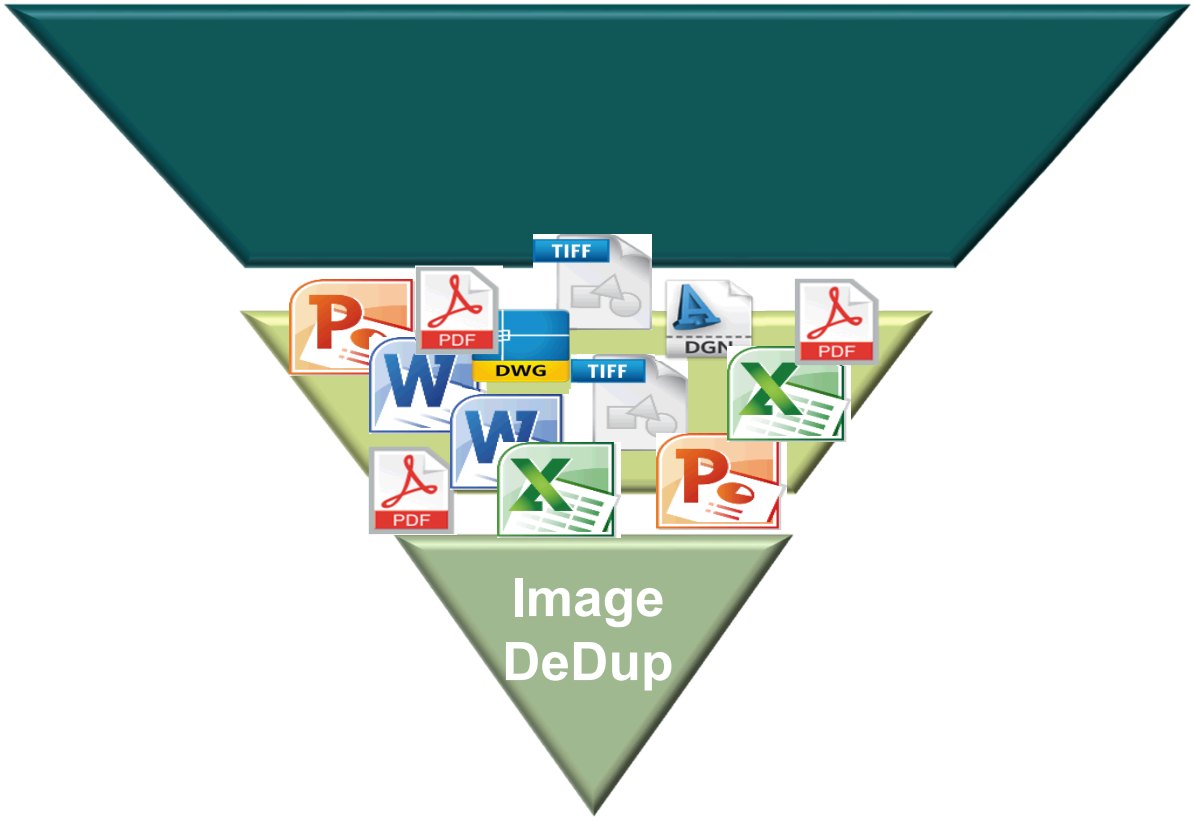
- XML extractions
- De-Duplication
- Auto classification
- Signature detection
- Contract comparison
- Revisions/versioning
- Expiration management
- Template confirmations




- Analyzes hash values of all files
- Duplicates identified & removed

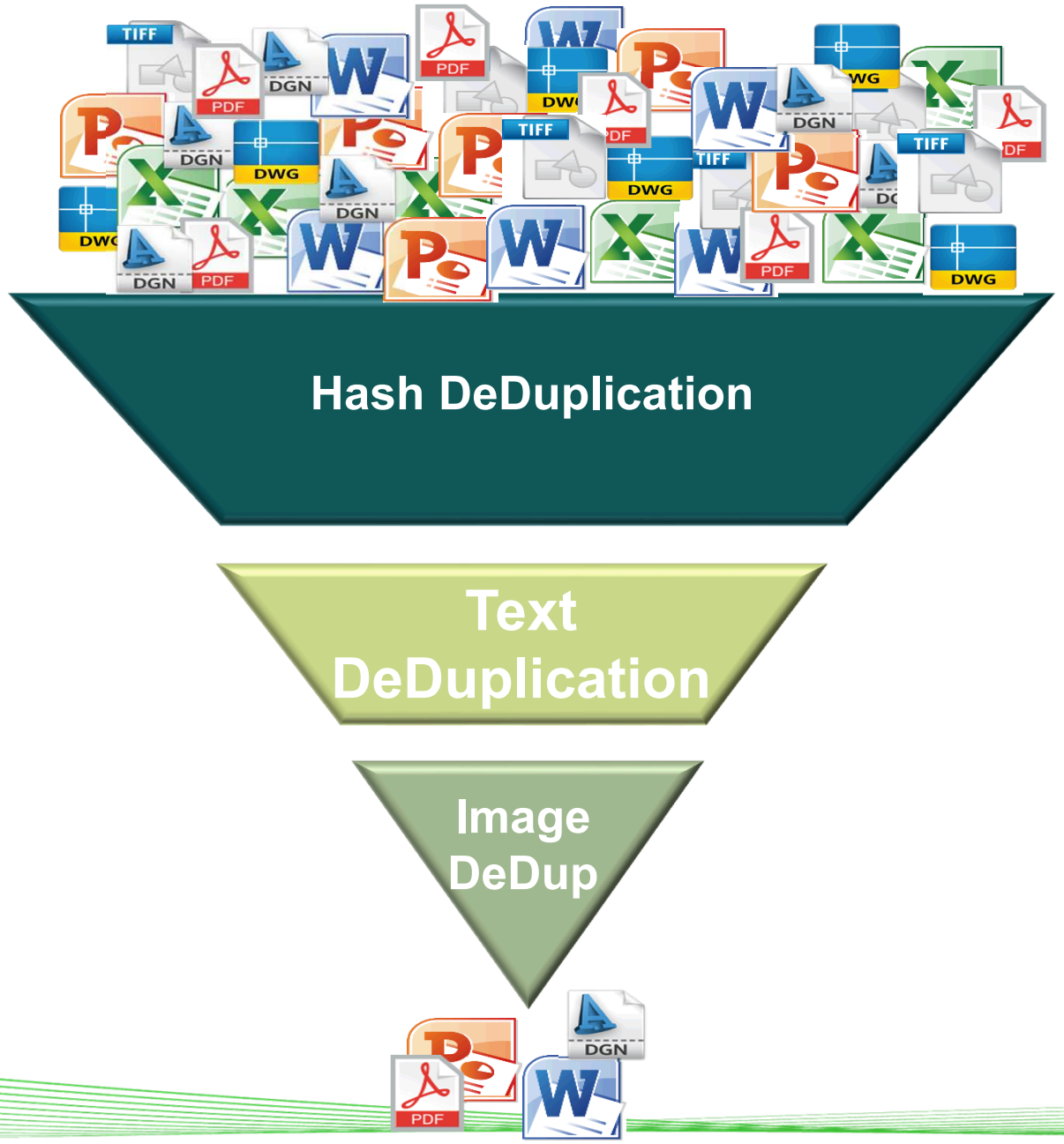



- Compares text (natively)
- OCR Image only content
- Duplicates identified based on threshold & removed




- Pixel by pixel comparison
- Duplicates identified based on threshold & removed

ADLIB®




# Powered By:

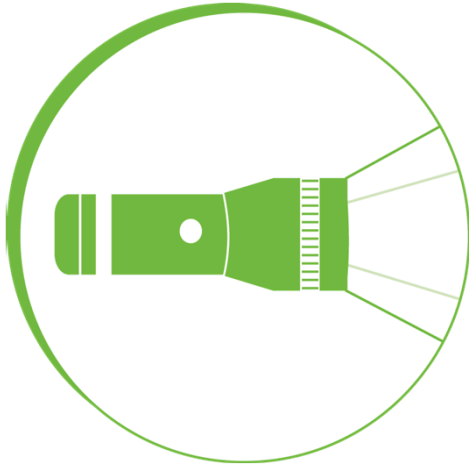
PDF

ADLIB

# Document Lifecycle Overview


Industry Expertize
PDF Technical Standard
Document Process Improvement

# Leveraging the PDF Standard to Understand Dark Data and Improve Document Processes

- Searchability
- Cost effective eDiscovery
- Classification/Deduplication
- Dirty Data
- Defensible Deletion
- Storage Optimization
- ROT reduction

ADLIB


Thank You
Questions:
www.adlibsoftware.com