OctoberPDFest
ONLINE

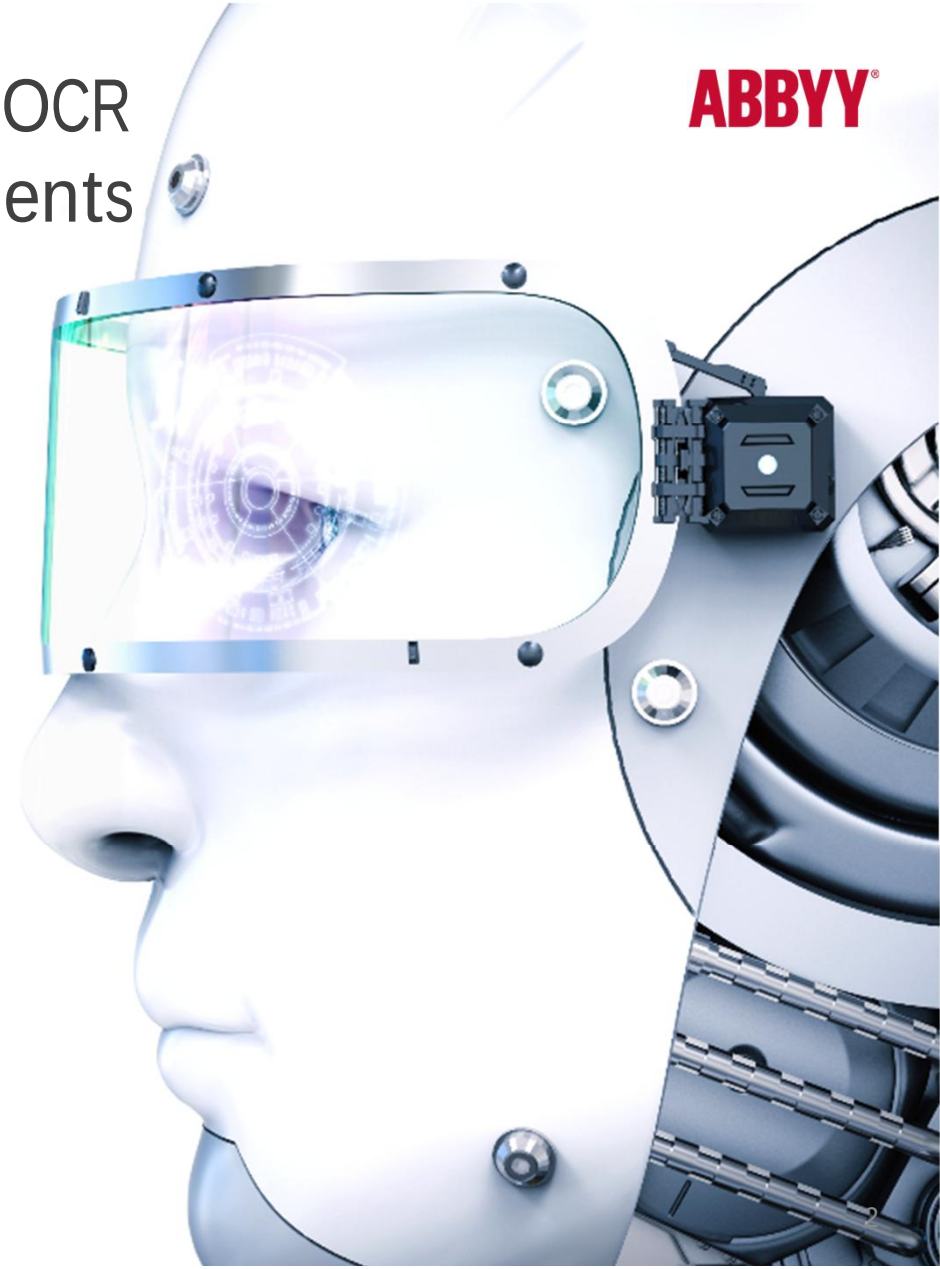
Digital Transformation and PDF

Why OCR is not only for capturing paper documents

Anna Koltsova
Director, Product Marketing, Mobile and FineReader, ABBYY Consumer Apps group

Digital Transformation and PDF – Why OCR Is Not Only for Capturing Paper Documents

- OCR – Simply Digitize?
- Why OCR for a PDF
- OCR Contribution
- Just Three of Many PDF features Made Possible by OCR
- FineReader PDF 15: The Smarter PDF Solution



OCR – Simply Digitize?

- OCR has been around for some time, so everybody knows what optical character recognition is for…
- It is for digitizing paper, isn't it?

à We would like to demonstrate that it does more than that. It allows to work with PDFs effectively.

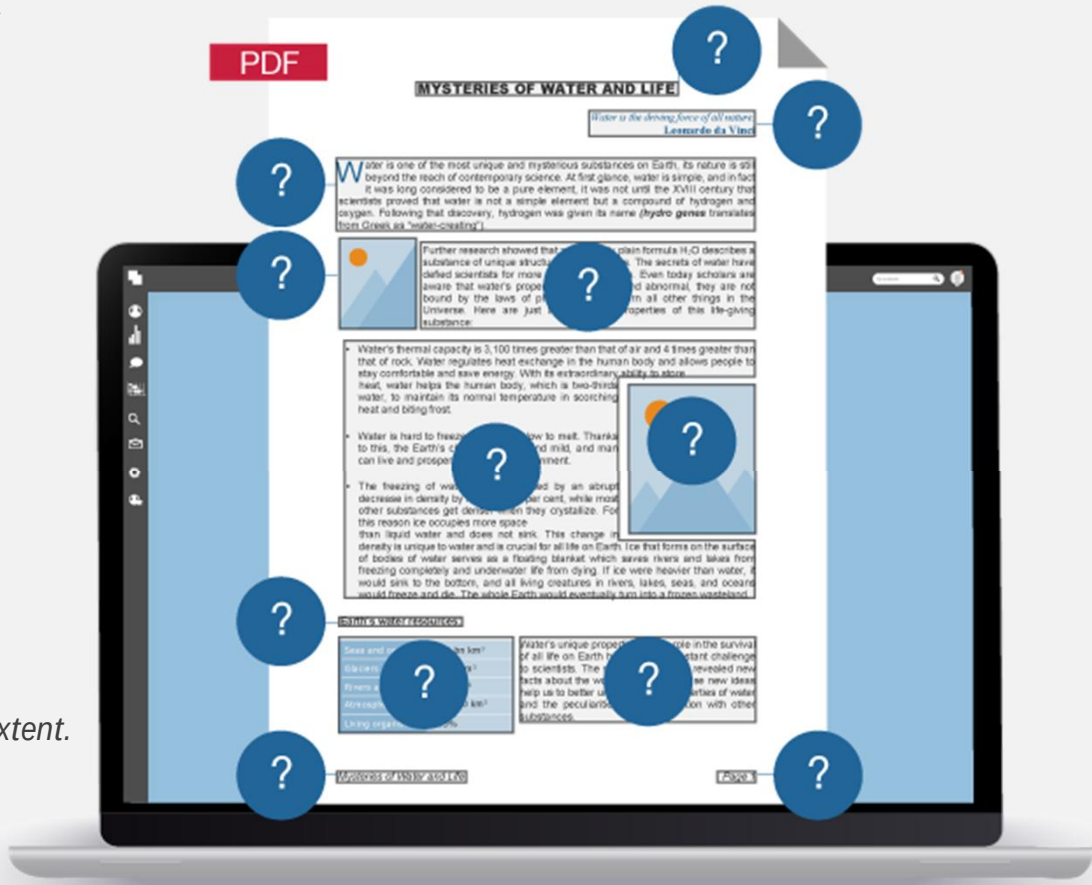
Why OCR for a PDF?

To modify or analyze a PDF, we need to know the document structure.

- Scanned "image-only" PDFs:
  - We don't have text to work with inside those
  - We don't know anything about their structure
- Searchable and "digital-born" PDFs:
  - We have text (not always)
  - We usually still don't have any information about the structure

PDFs do not contain information about the document structure*

** Some types of PDFs, such as PDF/UA, contain tags that describe it to some extent. Still, this may not be enough for preforming certain PDF operations.*



PDF Features Powered by OCR

ABBYY

Edit, Protect and Collaborate on PDFs

1. PDF Paragraph-level editing
2. URL autodetection and conversion into embedded links in scanned PDFs
3. Full-text search in scanned PDFs
4. Search & Redact in scanned PDFs
5. Extracting texts from scanned and problematic PDFs
6. Extracting tables from PDFs
7. Text mark-up in scanned PDFs

Create and Convert PDFs

8. MRC – effective PDF compression with minimal loss of visual quality
9. PreciseScan – effective improvement of visual quality of scanned documents
10. Saving to (creating) PDF/UA
11. Creation of Searchable PDFs
12. PDF conversion

Compare PDF Documents

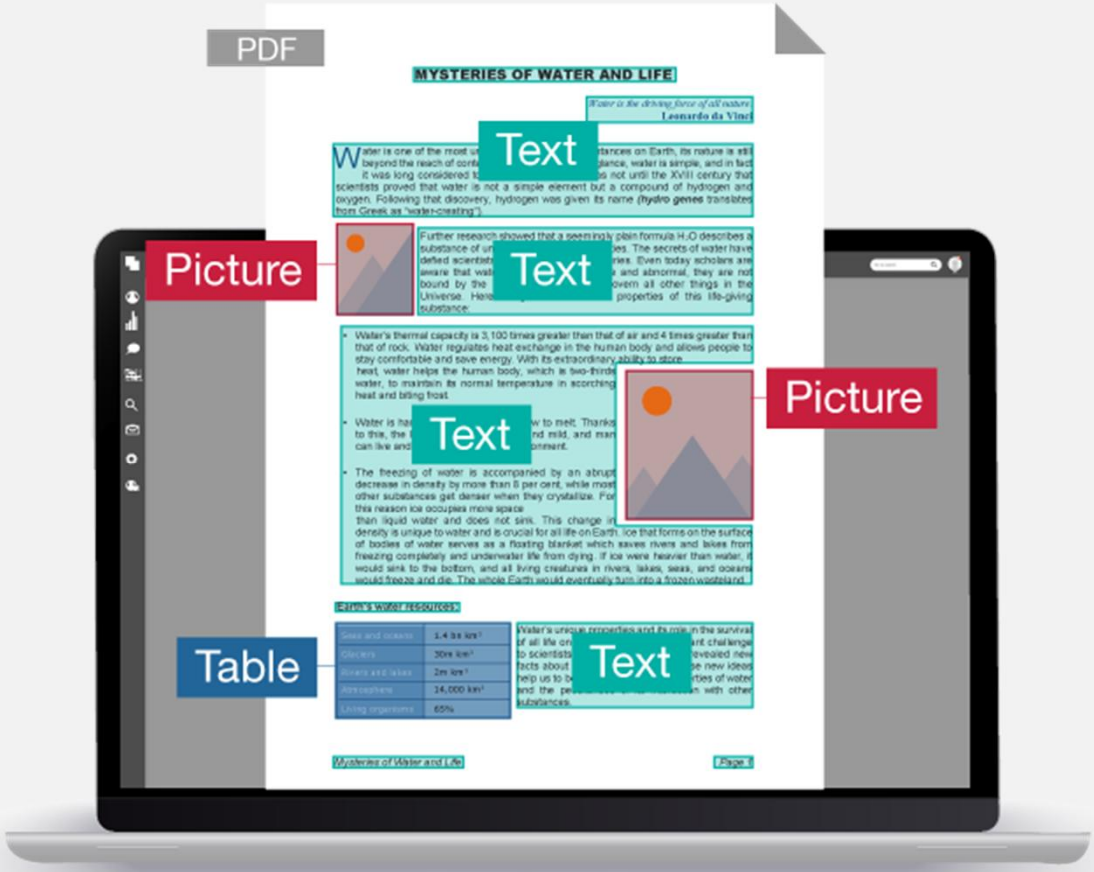
13. Comparing PDF documents

What Does OCR Do with a Document?

- Analyzes the pages – Document Analysis (DA) system

  DA finds:

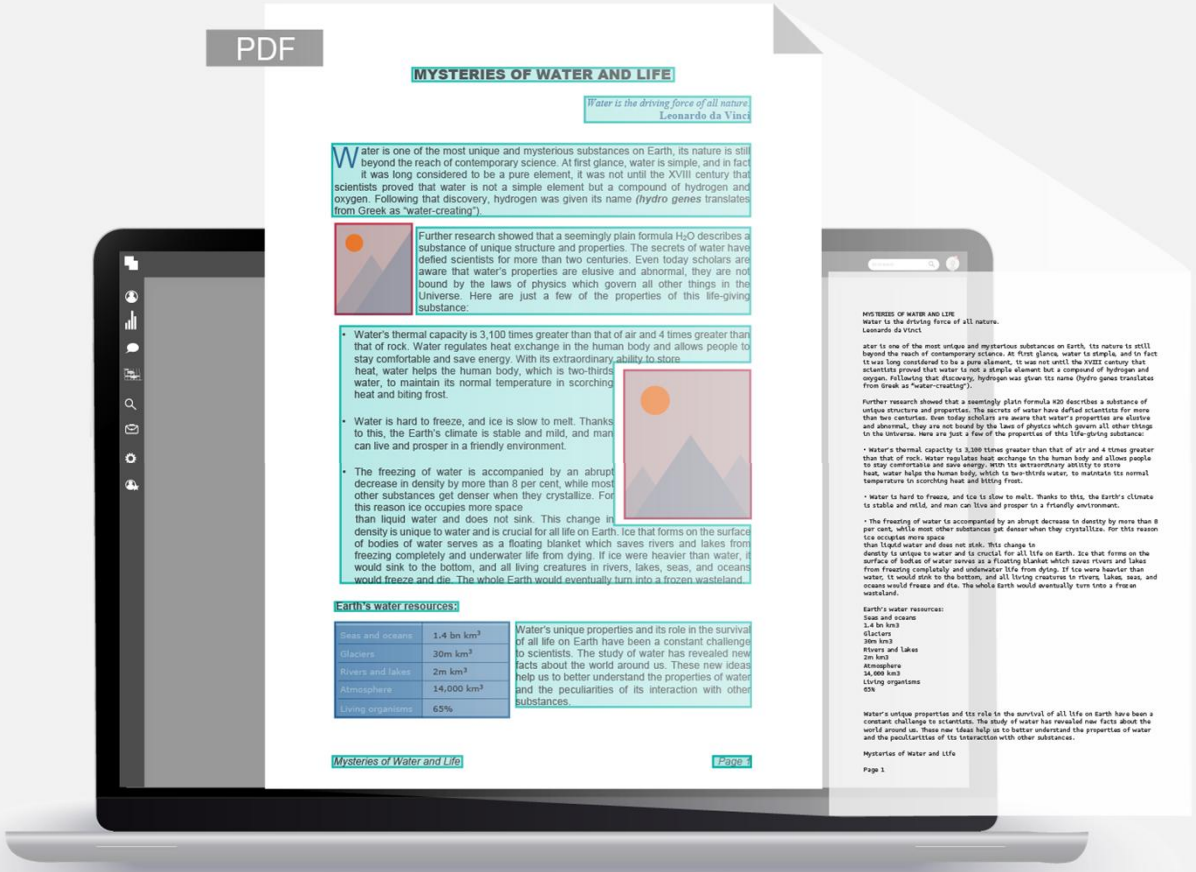
  - Texts, tables, pictures, background images, barcodes
  - Structure of tables: cells, separators and their types, background colors



What Does OCR Do with a Document?

- Recognizes the text – Optical Character Recognition (OCR) system itself

  OCR itself gives us digital text to work with

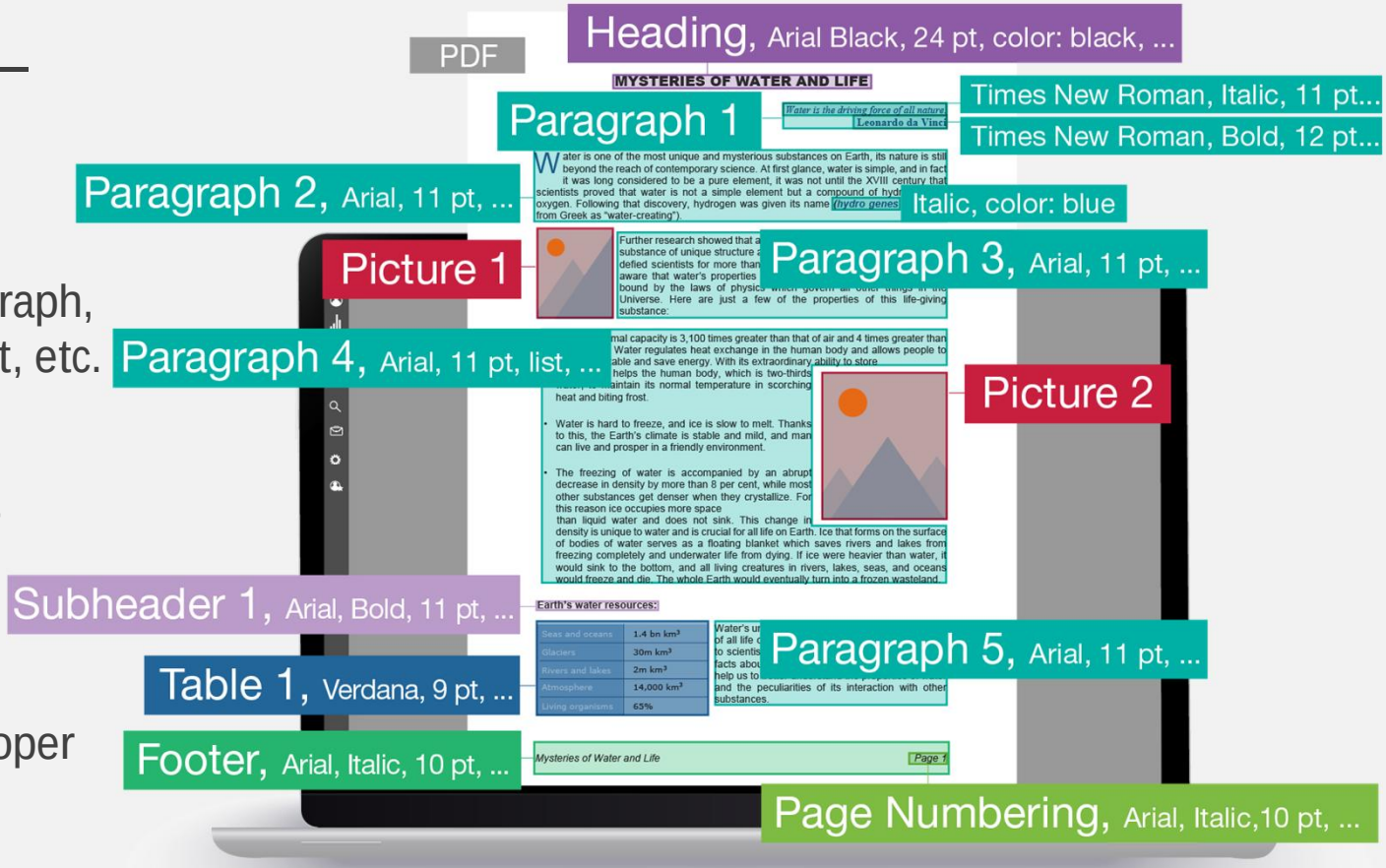


What Does OCR Do with a Document?

- Recreates the structure – Synthesis system

  Synthesis provides:

  - Roles of the text pieces: a paragraph, a heading, a header/footer, a list, etc.
  - Placement of these pieces on the page
  - Paragraph formatting: character and line spacing, indentations
  - Logically connected structure of a document, i.e., all its parts with information about their proper order and connections

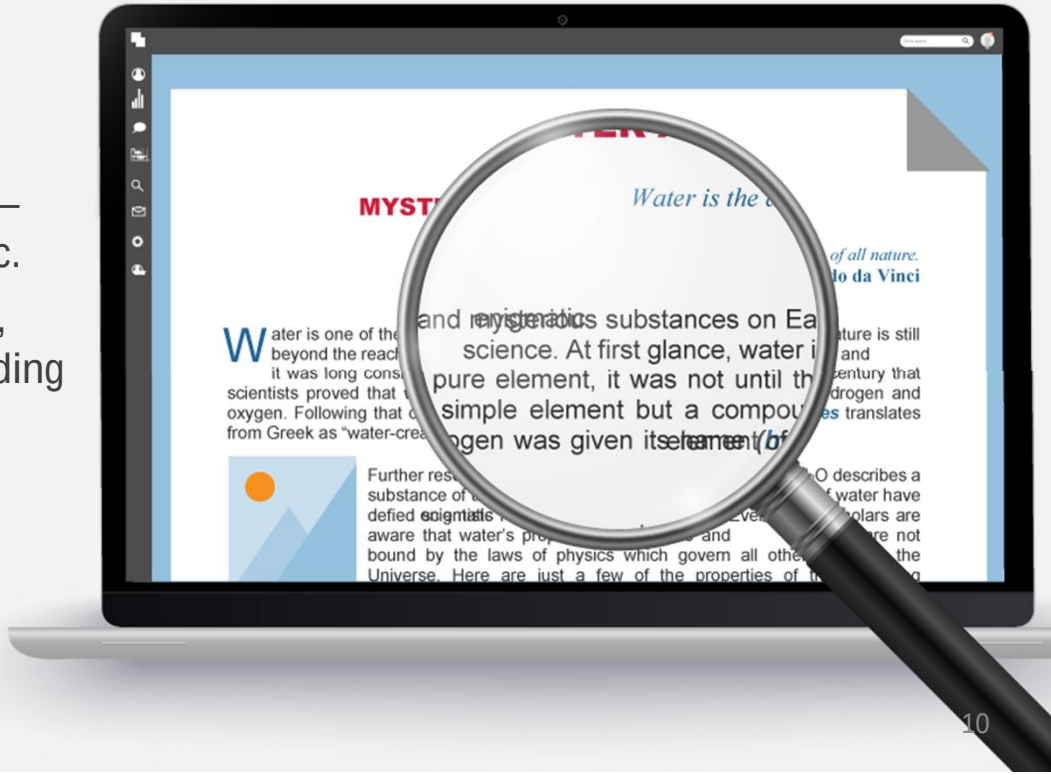


PDF Paragraph-Level Editing

PDF Paragraph-Level Editing – What's the Deal?

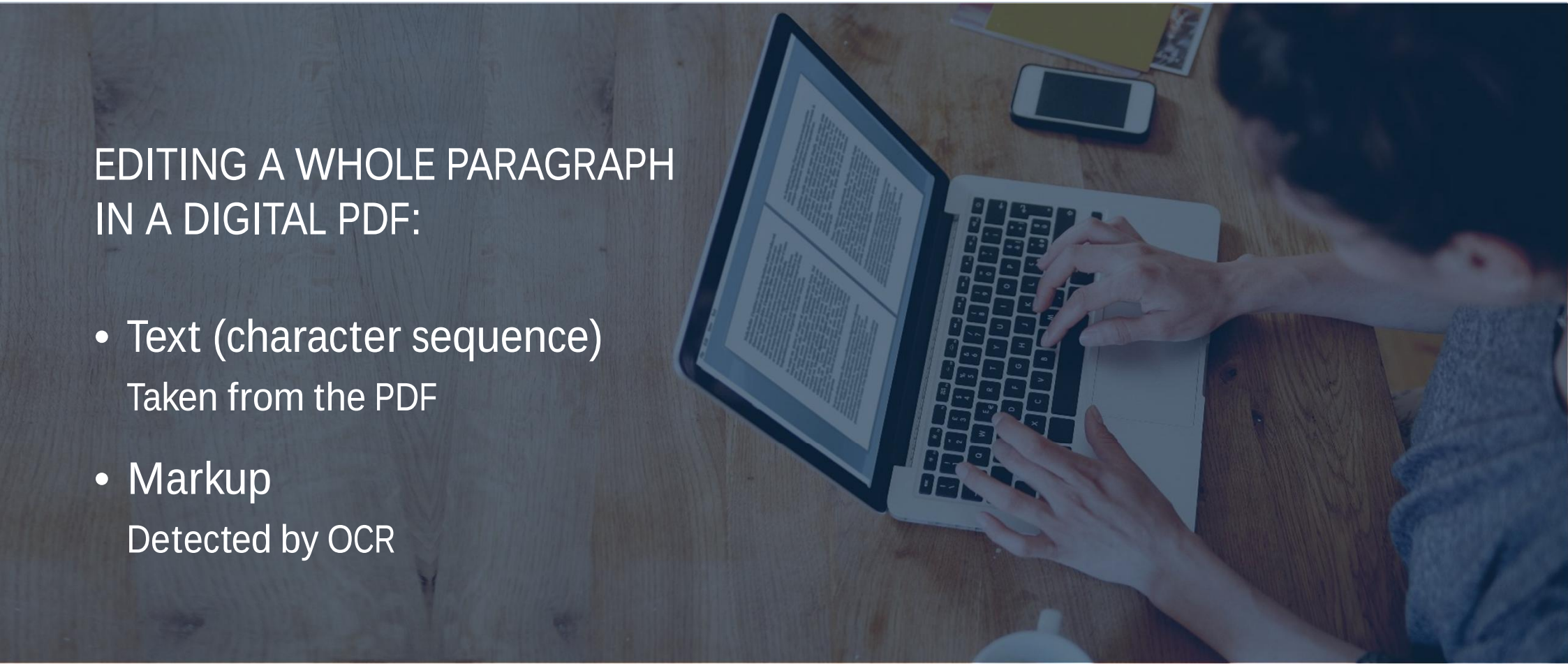
- Users' expectations: Digital PDFs should be like Word documents! There's text and everything else in them – it shouldn't be any problem to edit them.
- PDF reality:
  - PDF wasn't created with editability in mind. Rather otherwise. J
  - PDFs contain only information about separate characters – there's no information about words, lines, paragraphs, etc.
  - Simple adding or deleting of characters could be possible, but it wouldn't change the way and place where surrounding structurally connected characters are displayed.

There is no information in a PDF to know how to move other elements when adding or deleting some.



PDF Paragraph-Level Editing – OCR to the Rescue

ABBYY



EDITING A WHOLE PARAGRAPH IN A DIGITAL PDF:

- Text (character sequence)
  Taken from the PDF
- Markup
  Detected by OCR

PDF Paragraph-Level Editing – How It Is Done

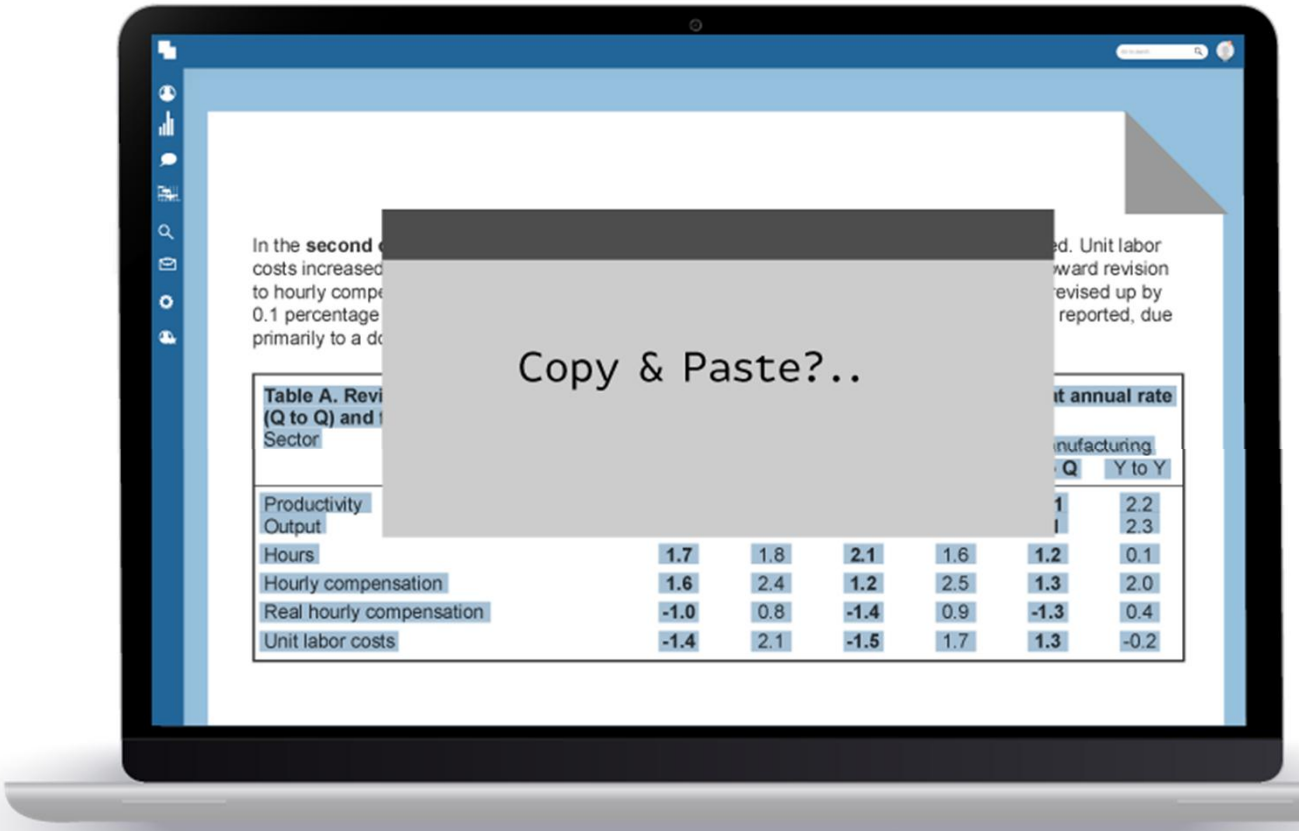
ABBYY

1	2	3	4	5
• DA processes raster image of the page and finds its elements	• Synthesis creates a temporary editable copy of the page with all necessary markup added	• Digital text from the PDF is aligned with the detected structure	• The user edits: text reflows from line to line; line and character spacing is followed; paragraph borders can expand or shrink according to the edits; and so on	• Once editing is done, the PDF is updated only in the part that has been changed. The rest remains original

Extracting PDF Tables

Extracting Tables: Can We Get What We Can See?

- We can copy text from a “digital-born” PDF (well, usually)
- ...so, can we copy tables the same way?



Extracting Tables: Why Not Just Copy?

ABBYY

We can’t copy the whole table:
we can just copy text from PDF tables, not more.

Visual appearance of a PDF table is defined by a set of objects (such as lines, rectangles) unrelated to the table content whatsoever.

- Why not copy and paste those objects? If we could do that, in the best case we could paste it only into another PDF. Still, success would be questionable.
- No way we could paste what we copied into Microsoft Word or Excel.

Extracting Tables: If We Can't "Read" about It in PDF, We Can "See" It!

ABBYY

!?

Relations of table elements in a digital-born PDF aren't described, so we can't "read" about tables in PDF

Let's "see" it: OCR can describe and recreate structure of a table based on its image

Extracting Tables: How It Is Done

ABBYY

1

- The selected area with a table is rasterized

2

- The image is analyzed by DA to find its elements:
  - Cells
  - Separators
  - Backgrounds
  - Text and picture elements in the cells

3

- The text is taken from the PDF (if the PDF text is good)
- If not, the text is also OCRed

4

- Synthesis “assembles” the detected structure and the content into a properly marked up piece, which a user can paste as a table into Excel, Word, etc.

Comparing PDF Documents

Comparing “Digitally-born” PDF Documents: We Have All the Characters, So What Else Do We Need?

ABBYY

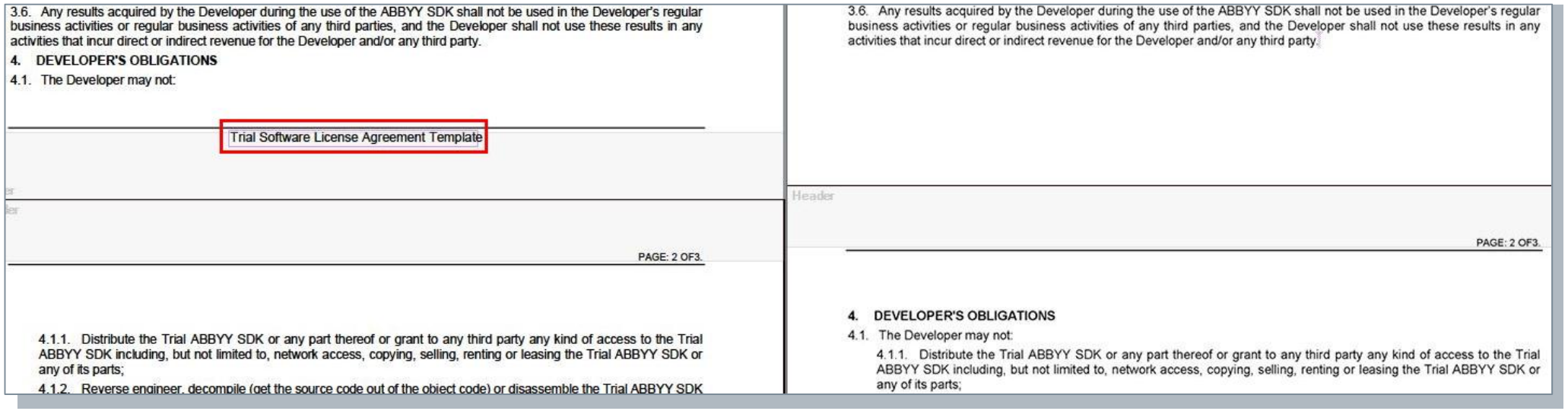
- When comparing two copies of a document, one of the goals is to minimize false differences.
- Main causes for false differences can be when, in the two copies:

Cause	What the solution is about
The same text is formatted in a different way or placed differently within the page	Having information about document structure
Main text is broken by a header/footer or an insert in different places	
OCR errors (if we must use OCR to get the text)	OCR accuracy

False Differences When the Structure Isn't Considered

ABBYY

One real difference – in the footer:



3.6. Any results acquired by the Developer during the use of the ABBYY SDK shall not be used in the Developer's regular business activities or regular business activities of any third parties, and the Developer shall not use these results in any activities that incur direct or indirect revenue for the Developer and/or any third party.

4. DEVELOPER'S OBLIGATIONS

4.1. The Developer may not:

Trial Software License Agreement Template

PAGE: 2 OF3.

4.1.1. Distribute the Trial ABBYY SDK or any part thereof or grant to any third party any kind of access to the Trial ABBYY SDK including, but not limited to, network access, copying, selling, renting or leasing the Trial ABBYY SDK or any of its parts;

4.1.2. Reverse engineer, decompile (get the source code out of the object code) or disassemble the Trial ABBYY SDK

3.6. Any results acquired by the Developer during the use of the ABBYY SDK shall not be used in the Developer's regular business activities or regular business activities of any third parties, and the Developer shall not use these results in any activities that incur direct or indirect revenue for the Developer and/or any third party.

Header

PAGE: 2 OF3.

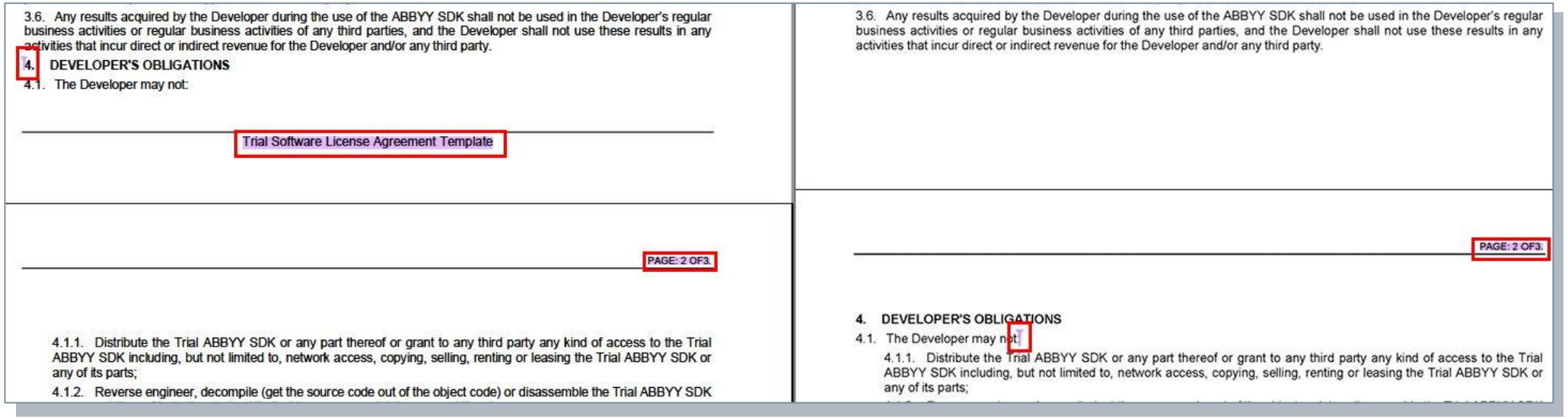
4. DEVELOPER'S OBLIGATIONS

4.1. The Developer may not:

4.1.1. Distribute the Trial ABBYY SDK or any part thereof or grant to any third party any kind of access to the Trial ABBYY SDK including, but not limited to, network access, copying, selling, renting or leasing the Trial ABBYY SDK or any of its parts;

False Differences When the Structure Isn't Considered

When comparison software doesn't know the structure:



3.6. Any results acquired by the Developer during the use of the ABBYY SDK shall not be used in the Developer's regular business activities or regular business activities of any third parties, and the Developer shall not use these results in any activities that incur direct or indirect revenue for the Developer and/or any third party.

4. DEVELOPER'S OBLIGATIONS

4.1. The Developer may not:

Trial Software License Agreement Template

PAGE: 2 OF3.

4.1.1. Distribute the Trial ABBYY SDK or any part thereof or grant to any third party any kind of access to the Trial ABBYY SDK including, but not limited to, network access, copying, selling, renting or leasing the Trial ABBYY SDK or any of its parts;

4.1.2. Reverse engineer, decompile (get the source code out of the object code) or disassemble the Trial ABBYY SDK

3.6. Any results acquired by the Developer during the use of the ABBYY SDK shall not be used in the Developer's regular business activities or regular business activities of any third parties, and the Developer shall not use these results in any activities that incur direct or indirect revenue for the Developer and/or any third party.

PAGE: 2 OF3.

4. DEVELOPER'S OBLIGATIONS

4.1. The Developer may not:

4.1.1. Distribute the Trial ABBYY SDK or any part thereof or grant to any third party any kind of access to the Trial ABBYY SDK including, but not limited to, network access, copying, selling, renting or leasing the Trial ABBYY SDK or any of its parts;

For one real difference,
four false differences / points of attention are created

Comparing “Digitally-born” PDF Documents: We Have All the Characters, So What Else Do We Need?

ABBYY

If we just take the text from a digital PDF and use it for comparison, we may create a lot of false differences.

Because we don’t know the document structure.

Plus, the text layer in such PDFs is not always accurate or usable.

Comparing “Digitally-born” PDF Documents – Is There Any Solution Good Enough?

ABBYY

WE SHALL USE OCR INTELLIGENTLY:

- Minimize utilization of character recognition…
- While getting and using enough information about the document structure

Comparing “Digitally-born” PDF Documents – How It Is Done

ABBYY

1

- Digital PDFs are “prepared”: rasterized and pre-recognized

2

- DA and Synthesis define the document structure:
  - Paragraphs
  - Lists
  - Inserts
  - Headers and footers
  - Page numbering, etc.

3

- Text layer quality analysis:
  - Valid digital text is taken from PDF
  - The bad parts are taken from OCR instead

4

- The texts are compared: paragraphs, headers, inserts from the two copies are now properly aligned – even if they do not pixel-match or even shifted in the two copies of a document

DEMO

MULTILINE PDF EDITING:

OCR recognizes text structure and allows editing of coherent text

EXTRACTING TABLES:

OCR can recognize tables so they can be copied and exported

COMPARING DOCUMENTS:

OCR kicks in if embedded text is bad and provides text to compare

White Paper (digital PDF).pdf - ABBYY FineReader PDF 15

File Edit View Tools Help

Rate This Product

New Task Open Save As 10 / 11 Add Pages Rotate Crop Recognize Compare Form Editor PDF Tools Comments (0)

PAGES

Note Draw Edit Erase Picture Watermark Add Stamp Header and Footer Signature Redact Password Security

Add Text


8


9


10


11

Definitions

Computer vision: focused on the automatic extraction, analysis, and understanding of information from images, including scanned documents.

Machine Learning: refers to software that enables machines to "learn" in both real-time or over time, improving accuracy and performance. In a process involving capturing documents and processing with RPA, machine learning and other AI technology learns from potentially thousands of variations of documents such as processing invoices or handling vendor orders.

Natural Language Processing: The understanding of human lan-

AI (Artificial Intelligence) has become a huge buzzword in today's business and can be a lot of different things to users and buyers of technology. Until now, most companies have been myopic when it comes to AI––not fully understanding the best way to use it. In many cases organizations have failed to see AI's full potential and settled for only incremental improvements to traditional processes. To maximize AI's impact, organizations must understand how AI is applied to content and how when coupled with RPA it can create new, self-generating, self-optimizing processes that learn "as-they-go".

The big difference between RPA and AI technologies is RPA is focused on repetitive structured work while AI technologies are designed to understand unstructured content. AI applies intuition, learned judgment and problem solving pertaining to the process and content associated with it. RPA works with only structured data, which represents only a subset of processes that organizations are looking to automate.

"AI technology – including Machine Learning (ML), Natural Language Processing (NLP) and computer vision – used in combination with RPA has the potential to unlock significantly greater business benefits than that achievable through standalone RPA. The combination can help automate not just the transactional portions of processes but also the judgement-intensive ones."

– Anil Vijayan, Practice Director, Service Optimization Technologies

Background recognition is in progress... 1:1 200%


ABBYY® FineReader® PDF 15
The smarter PDF solution

FineReader PDF 15‘s Unique OCR Benefits

ABBYY

Excellent OCR results through exceptional AI-based ABBYY technology

Ease of use because of unique background recognition

Flexibility of PDF content usage thanks to layout recognition

Copying Tables with FineReader PDF

We love ABBYY Fine Reader. We receive price lists in a PDF format which we formerly had to copy and paste pieces of information to Excel. One price list took two of us at least a day and a half to process the data to Excel. With ABBYY we were able to convert the same document to Excel in under 5 minutes! The time we are saving more than compensates for the price of the software. The ROI was immediate!

Gloria Coleman,
Consultant, Smartwyre

Making Content Accessible with FineReader PDF

ABBYY

As an Alternative Media Specialist, I use ABBYY FineReader to enhance the accessibility and quality of PDF documents for students and educators. PDF documents sometimes pose challenges; FineReader is my go-to tool for generating a searchable, accessible or alternative format. When processing alternative media requests, FineReader is always a part of my workflow when processing PDF documents.

I am legally blind and I also use FineReader to produce documents that work with my text-to-speech and screen reading technology. My favorite feature FineReader offers is the ability to convert an inaccessible PDF into a universally accessible document or PDF/UA.

Mathew Spinneberg,
Ventura County Community College

The FineReader Product Family

ABBYY

Individual productivity

FINEREADER PDF STANDARD

- Daily work with PDF documents
- Document collaboration
- Conversion and reusing content
- PDF creation

Workgroup productivity

FINEREADER PDF CORPORATE

- Daily work with PDF documents
- Document collaboration
- Conversion and reusing content
- PDF creation
- Document comparison
- Automating and scheduling conversion

Document conversion as a service

FINEREADER SERVER

- Digital archiving, long-term storage, and compliance
- Document digitization for further processing, OCR, and file conversion service
- Centralized automated document conversion service for all employees
- Custom integration into other business systems
- Indexing and automated document separation

Can be combined with FineReader Server

Can be combined with FineReader Standard or Corporate

Get FineReader Corporate

ABBYY

Promocode **OctoberPDFest20**

- **20%discount** on FineReader Win Corporate Edition.
- Valid until 31 December 2020.
- 3 units per order.

PDF.ABBYY.COM

ABBYY

PDF.ABBYY.COM

Thank you!

anna.koltsova@abbyy.com

LinkedIn: Anna-Koltsova