



Robotic Process Automation (RPA) and PDF

Robo-PDF Apocalypse or Automated PDF Utopia?

Version - 8 Sept 2020

Neil Pitman
Managing Director, Aquaforest Limited

Monday, October 27th 2020 5PM CET

OctoberPDFest
ONLINE

Session Overview

- Robotic Process Automation (RPA)
 - Uses software robots to mimic human actions and automate everyday tasks.

- This session will cover
 - The place that PDF plays in the RPA
 - How RPA fits in the wider spectrum of automation.
 - Typical use cases
 - PDF-centric issues that need to be considered for successful implementations.

Agenda

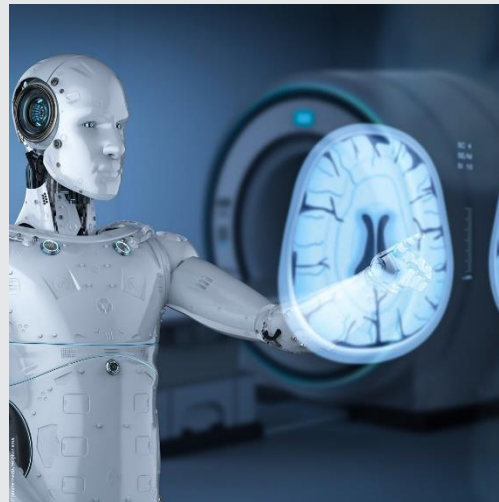


- RPA and the Automation Spectrum
- Demonstration : PDF and RPA
- PDF and RPA : The Challenge

RPA and the Automation Spectrum



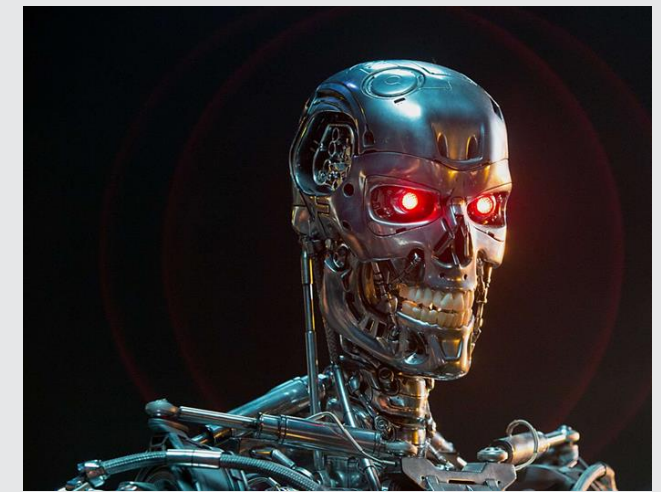
**Robotic process
Automation**



**Intelligent
Automation**



**Cognitive
Intelligence**



**Artificial
Intelligence**

Robotic Process Automation

- Mimics Human Actions
- Used for rules based processes
- Enables
 - Faster handing time
 - Higher volumes
 - Reduced errors & costs
 - Without impacting existing systems



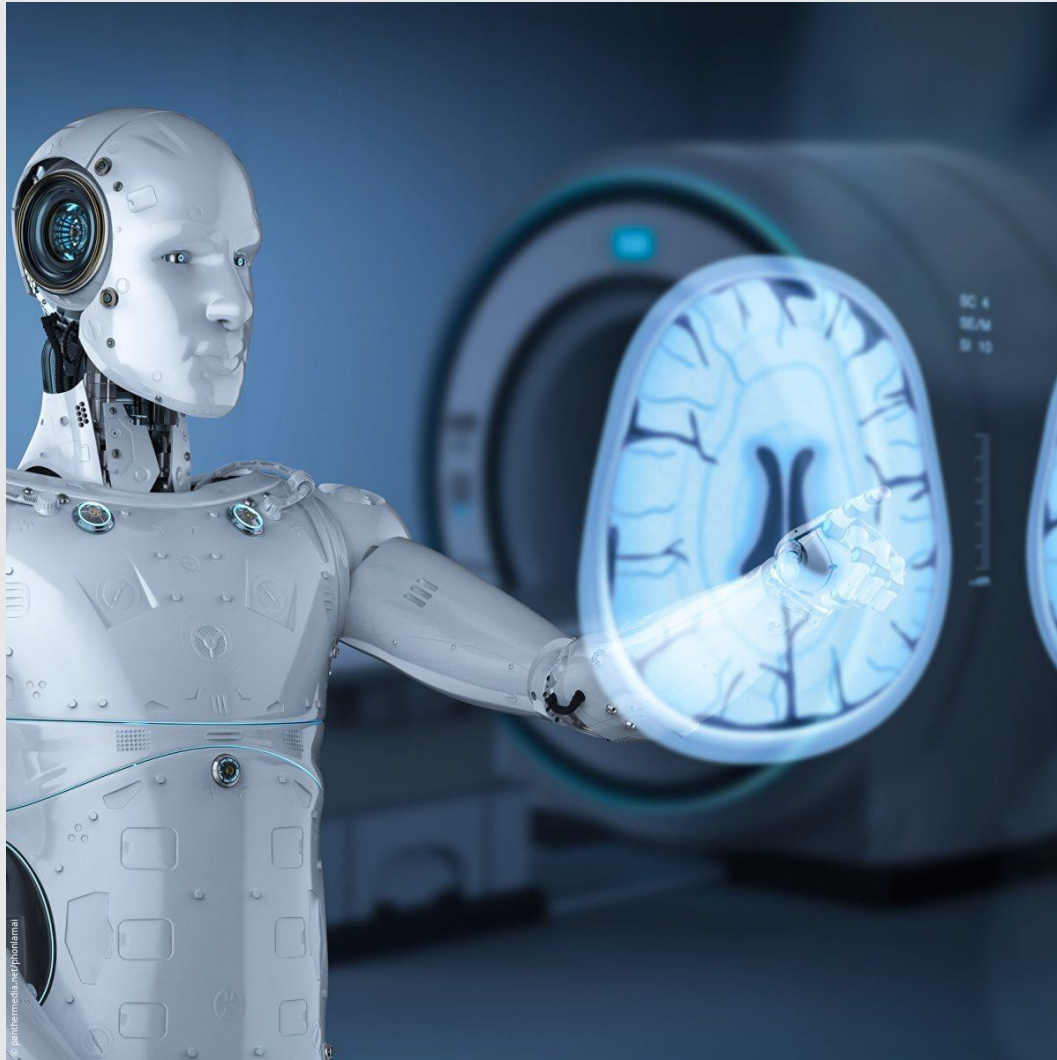
Typical Processes

- Business Process
 - Consistent and routine
 - High volume
 - Prone to human error
 - Limited in requiring a person to make a decision
 - Manual data entry or “swivel chair” work
- Technology
 - Legacy Systems
 - No APIs

Common Use Cases

- Payroll
- Employee status changes
- New hire recruitment and onboarding
- Accounts receivable and payable
- Invoice processing
- Inventory management
- Report creation
- Updating CRM data
- Software installations
- Data migration
- Vendor onboarding

Intelligent Automation

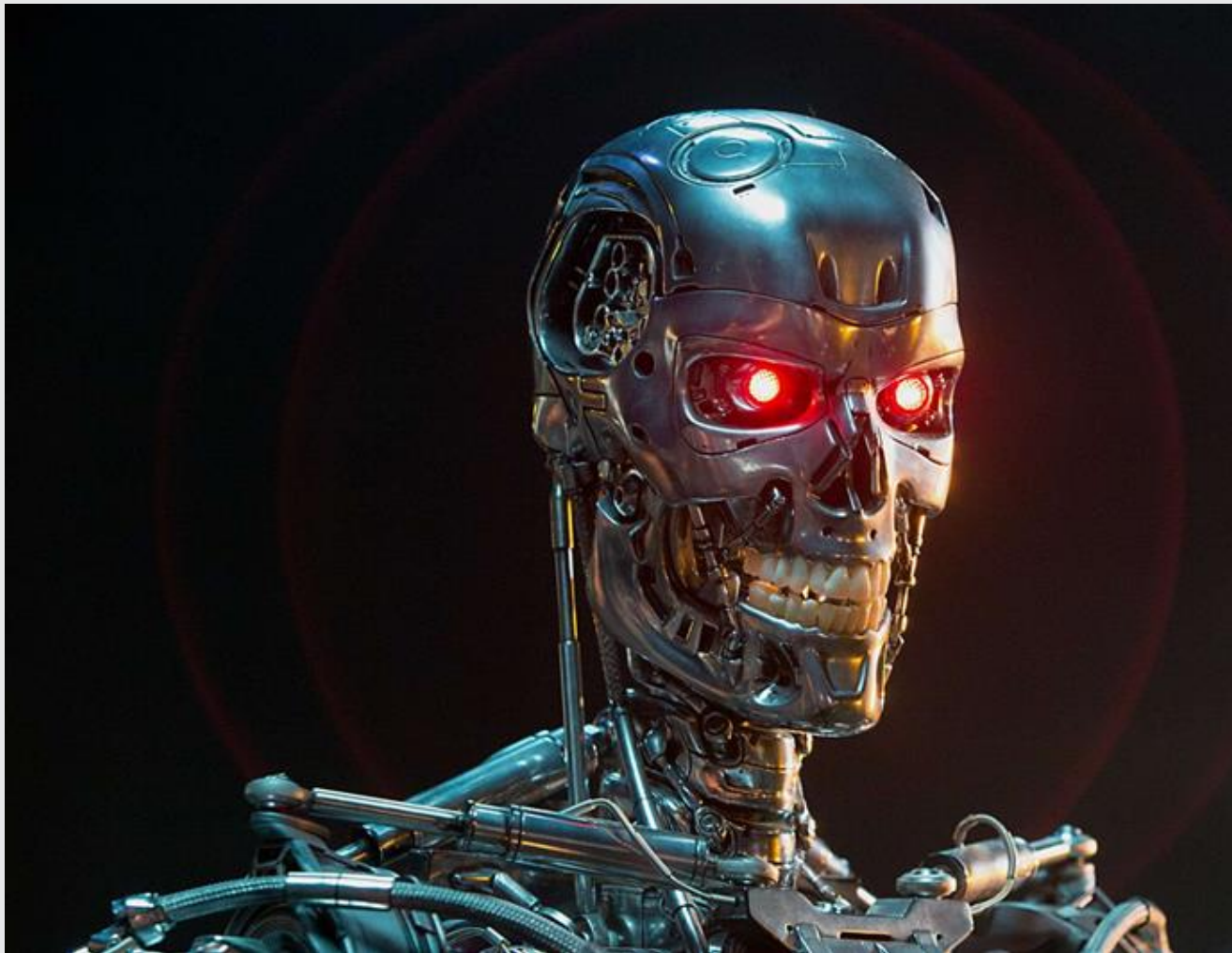


- Mimics Human Judgement
- Used for judgement based processes
- Machine learning capability
- Interprets human behaviour

Cognitive Intelligence

- Imitate Human Processes
- Natural Language
- Image Recognition
- Augments Human Intelligence

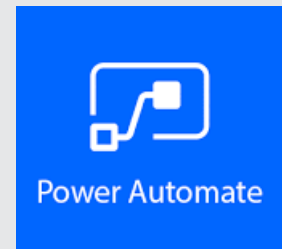




- Mimics Human Intelligence
- Acquires human-like thought processing and thinking capabilities

RPA Demonstration

- Demonstration using Microsoft Power Automate



- Many other RPA platforms are available



Asset Registration Use Case

Asset Registration

Asset Form PDF Days-2.pdf - Adobe Acrobat Pro

File Edit View Plug-Ins Window WindJack Tools Help

Open Create Save Print Mail Comment Fill & Sign

1 / 1 100%

New Asset Registration Form

2020 Edition Reference AB78

ID	75" Monitor	Category	Hardware
Condition	As New	Location	Berlin
Purchase Price	634.00	Current Value	500.00
Owner	Aquaforest	Acquired Date	1/1/2020
Approved By	J. K. Rowlino		

Database2 : Database- C:\Users\Neil\Desktop\Database2.accdb (Access 2007 - 2016 file format) - Access

Asset Details

Save and New E-mail Print Close

Item

Asset Details Comments

Category	Hardware
Manufacturer	
Model	
Acquired Date	
Purchase Price	\$634.00
Current Value	\$500.00
Condition	As New
Location	Berlin
Owner	Aquaforest
Retired Date	

(attachment preview)

Add or Remove Attachments

RPA Workflow



Aquaforest Power Automate

Search for helpful resources

Environments Neil Pitman's Environm... ? NP

PDF Days • Ran at 02/03/2020 14:19:52

Resubmit Cancel Edit

✓ Your flow ran successfully.

When a file is created 1s ✓

↓

Get data from PDF 35s ✓

↓

Run a UI flow for desktop 2m ✓

Get The PDF Data



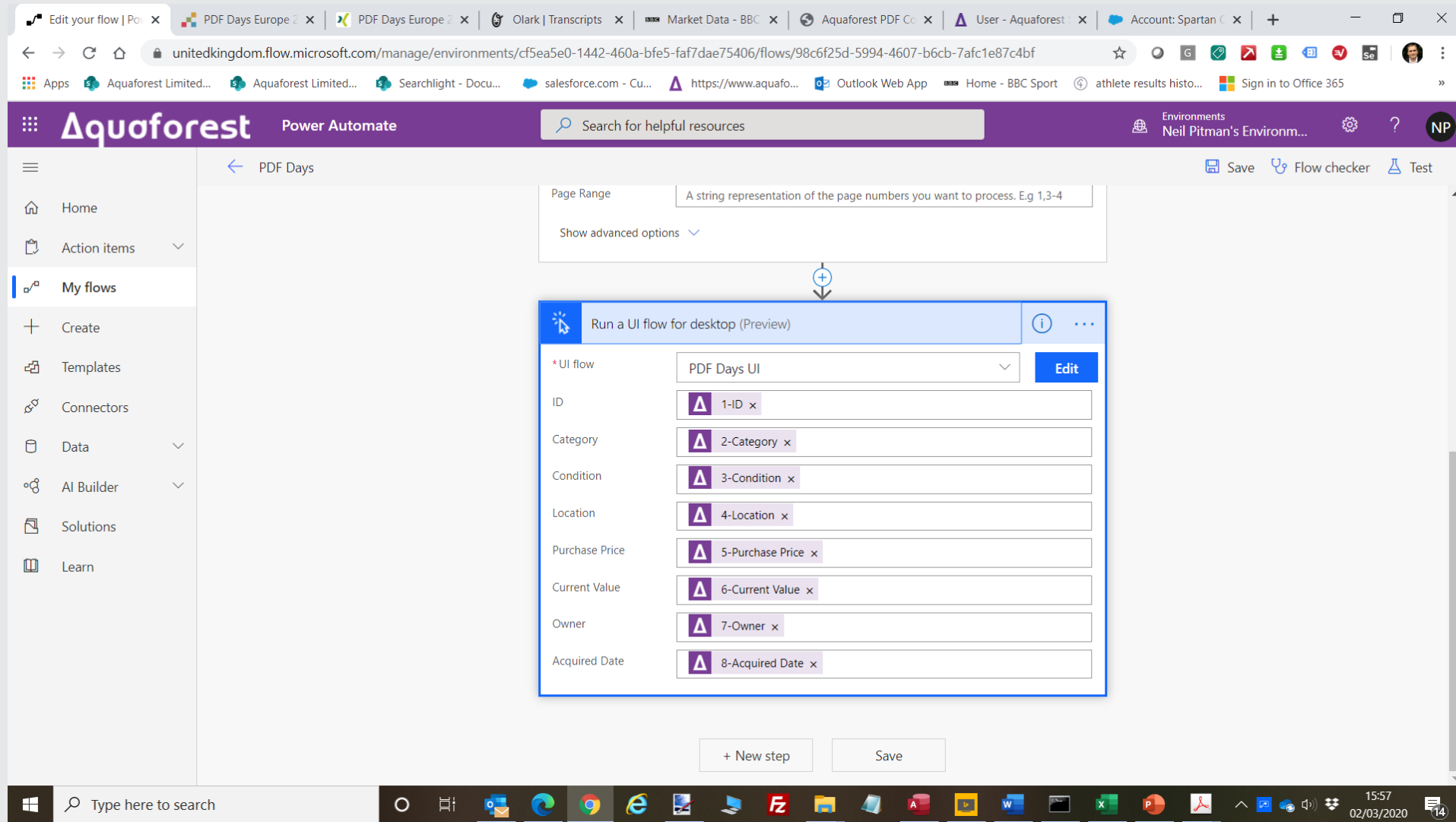
Screenshot of the Aquaforest Power Automate interface showing a flow named 'PDF Days'.

The flow consists of two steps:

- When a file is created**: Triggered by a file being created in the folder `/Flow2/RPA`.
- Get data from PDF**: Extracts data from the PDF file. The 'File Content' is set to 'File content'. The 'Expected Keys' are listed as: ID, Category, Condition, Location, Purchase Price, Current Value, Owner, and Acquired Date. The 'Page Range' is set to 'A string representation of the page numbers you want to process. E.g 1,3-4'.

The interface includes a sidebar with navigation options (Home, Action items, My flows, Create, Templates, Connectors, Data, AI Builder, Solutions, Learn) and a top navigation bar with the Aquaforest logo and search bar.

Send The Data to the UI Flow



The screenshot displays the Microsoft Power Automate web interface. The top navigation bar includes the 'Aquaforest' logo, 'Power Automate' text, a search bar, and user information. The left sidebar shows navigation options: Home, Action items, My flows (selected), Create, Templates, Connectors, Data, AI Builder, Solutions, and Learn. The main workspace is titled 'PDF Days' and contains a 'Page Range' input field with the value 'A string representation of the page numbers you want to process. E.g 1,3-4'. Below this, a 'Show advanced options' dropdown is visible. A modal window titled 'Run a UI flow for desktop (Preview)' is open, showing a list of UI flow steps: 'PDF Days UI' (selected), '1-ID', '2-Category', '3-Condition', '4-Location', '5-Purchase Price', '6-Current Value', '7-Owner', and '8-Acquired Date'. Each step has a small 'x' icon to its right. The modal also includes an 'Edit' button. At the bottom of the modal, there are '+ New step' and 'Save' buttons. The bottom of the screen shows the Windows taskbar with various application icons and the system clock indicating 15:57 on 02/03/2020.

UI Flow Automates User Actions



PDF Days UI | Forum | Save | Close

Record and edit steps

Use your mouse and keyboard to recreate the steps of the process you want to automate. [Learn more](#)

Start your UI flow

Run Access script

- Launch Access
- LeftClick 1
- LeftClick 2

Back Next

Database2 : Database- C:\Users\Neil\Desktop\Database2.accdb (Access 2007 - 2016 file format) - Access

Asset Details

Save and New | E-mail | Print | Close

Item 75" Monitor

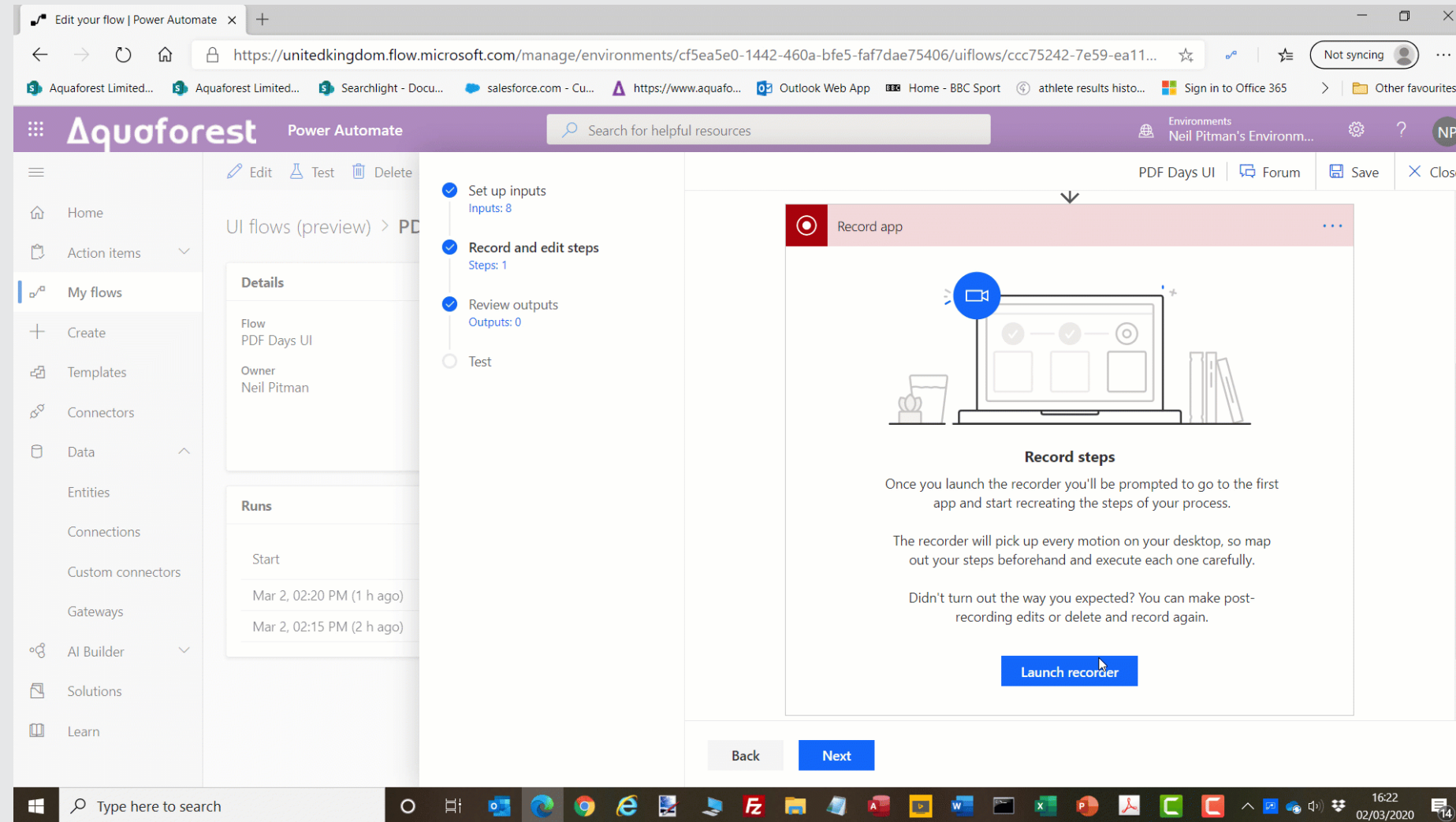
Asset Details | Comments

Category	Hardware
Manufacturer	
Model	
Acquired Date	
Purchase Price	\$634.00
Current Value	\$500.00
Condition	As New
Location	Berlin
Owner	Aquaforest
Retired Date	

(attachment preview)

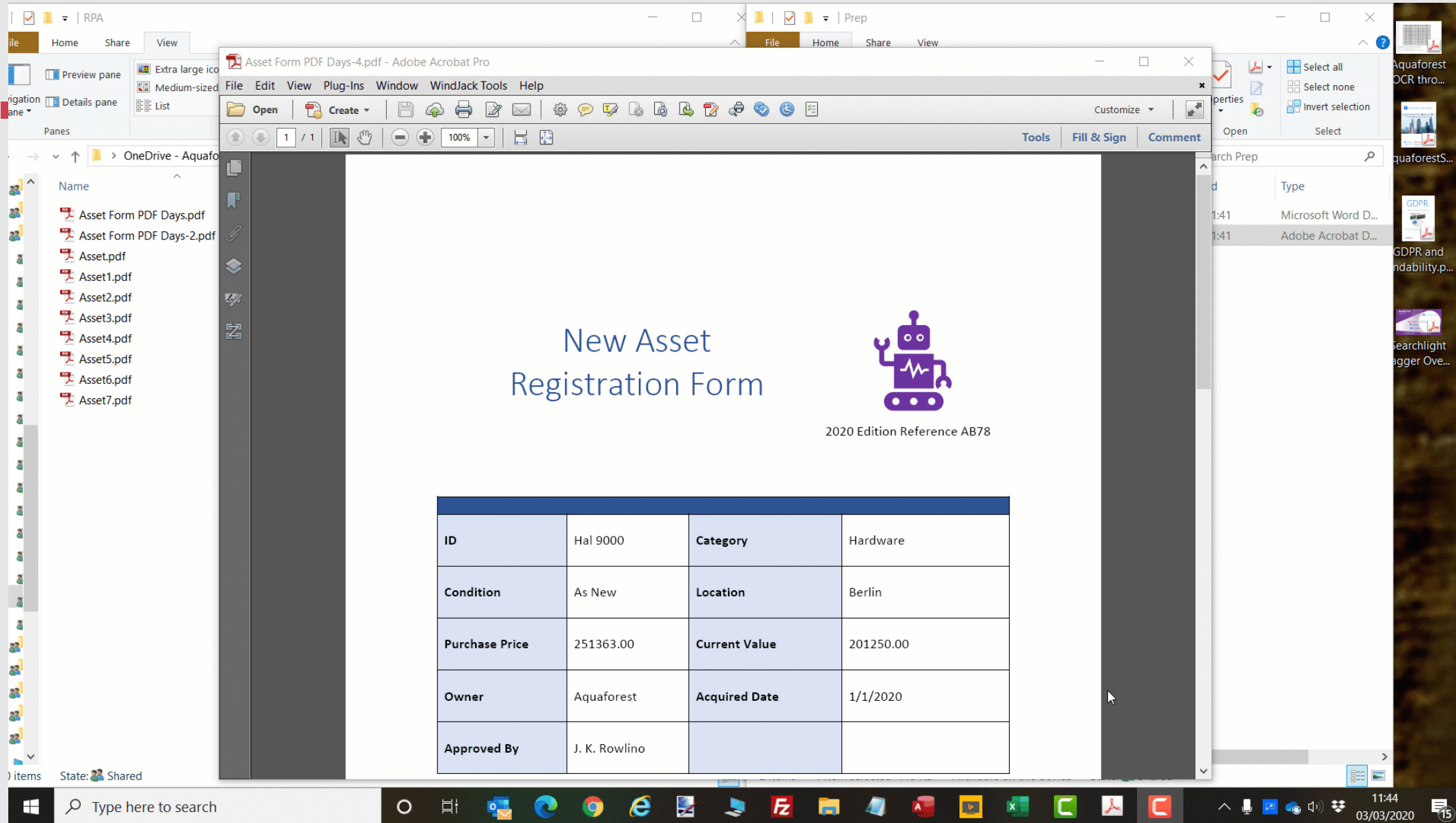
Add or Remove Attachments

Demo Video#1 – Recording UI Flow



The screenshot displays the Microsoft Power Automate web interface. The browser address bar shows the URL: <https://unitedkingdom.flow.microsoft.com/manage/environments/cf5ea5e0-1442-460a-bfe5-faf7dae75406/uiflows/cc775242-7e59-ea11...>. The interface includes a left-hand navigation pane with options like Home, Action items, My flows, Create, Templates, Connectors, Data, Entities, Connections, Custom connectors, Gateways, AI Builder, Solutions, and Learn. The main area shows a flow named 'PDF Days UI' with a 'Record app' step selected. The 'Record app' step is detailed with a 'Record steps' section that explains the recording process: 'Once you launch the recorder you'll be prompted to go to the first app and start recreating the steps of your process. The recorder will pick up every motion on your desktop, so map out your steps beforehand and execute each one carefully. Didn't turn out the way you expected? You can make post-recording edits or delete and record again.' A 'Launch recorder' button is visible at the bottom of the step details. The top right of the interface shows 'Not syncing' and a user profile icon. The bottom of the screen shows the Windows taskbar with various application icons and the system clock indicating 16:22 on 02/03/2020.

Demo Video#2 – RPA Running



The screenshot shows a Windows desktop environment. In the foreground, a window titled "Asset Form PDF Days-4.pdf - Adobe Acrobat Pro" is open. The document displays a "New Asset Registration Form" with a purple robot icon and the text "2020 Edition Reference AB78". Below the form, a table contains the following data:

ID	Hal 9000	Category	Hardware
Condition	As New	Location	Berlin
Purchase Price	251363.00	Current Value	201250.00
Owner	Aquaforest	Acquired Date	1/1/2020
Approved By	J. K. Rowllino		

The background shows a file explorer window with a list of PDF files in OneDrive, including "Asset Form PDF Days.pdf", "Asset Form PDF Days-2.pdf", and various "Asset" files. The Windows taskbar at the bottom shows the time as 11:44 on 03/03/2020.

“Unassisted” Use Case

Yet Another Asset - Message (HTML)

File Message Insert Options Format Text Review Help Adobe PDF Tell me what you want to do

Clipboard: Cut, Copy, Paste, Format Painter

Basic Text: Font size (11), Bold (B), Italic (I), Underline (U), Text color, Background color, Bulleted list, Numbered list, Indent, Outdent

Names: Address Book, Check Names

Include: Attach File, Attach Signature, Item

Tags: Follow Up, High Importance, Low Importance

Add-ins: Office Add-ins, Dictation

Send

From: neil.pitman@aquaforest.com

To: Neil Pitman

Cc:

Subject: Yet Another Asset

Asset Form PDF Days-4.pdf 114 KB

Best Regards

Neil Pitman
Aquaforest

Aquaforest
Making Content Findable
Over 2000 Organizations in 50+ Countries rely on Aquaforest
Email: info@aquaforest.com Web: www.aquaforest.com
AQUAFOREST LTD - ENGLAND - COMPANY NO: 4344383
REGISTERED OFFICE: MIDSHIRES HOUSE, SUITE 32, MIDSHIRES BUSINESS PARK, LUTON, LU1 3AB

Asset Form PDF Days-4.pdf - Adobe Acrobat Pro

File Edit View Plug-Ins Window WindJack Tools Help

Open Create Save Print Comment Tools Fill & Sign

1 / 1 56.2%

New Asset Registration Form

2020 Edition Reference AB78

ID	Hal 9000	Category	Hardware
Condition	As New	Location	Berlin
Purchase Price	251363.00	Current Value	201250.00
Owner	Aquaforest	Acquired Date	1/1/2020
Approved By	J. K. Rowlinson		

Apps Aquaforest Limited... Aquaforest Limited... Searchlight - Docu... salesforce.com - Cu... https://www.aquafo...

Excel Neil Pitman > PDF Days

Table Tools PDF Days Asset Register

File Home Insert Formulas Data Review View Help Design Tell me what you want to do

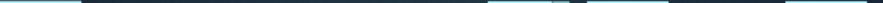
Undo Paste Cut Copy Format Painter

Clipboard Font Alignment Number

A6 Hal 9000

	A	B	C	D	E	F	G
1	Asset List						
2							
3	Description	Category	Condition	Location	Purchase Price	Current Value	Owner
4	Paper Shredder	Hardware	As New	Berlin	634	500	Aquaforest
5	75" Monitor	Hardware	As New	Berlin	634	500	Aquaforest
6	Hal 9000	Hardware	As New	Berlin	251363	201250	Aquaforest
7							
8							
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							
19							
20							
21							
22							

A horizontal beam is shown with a downward-pointing arrow at its center. The arrow is labeled with a plus sign (+) inside a circle, indicating a positive shear force.



Simple Unassisted Case

Yet Another Asset - Message (HTML)

File Message Insert Options Format Text Review Help Adobe PDF Tell me what you want to do

Clipboard Basic Text Names Include Tags Add-ins Voice

From: neil.pitman@aquaforest.com

To: Neil Pitman

Cc:

Subject: Yet Another Asset

Asset Form PDF Days-4.pdf 114 KB

Best Regards

Neil Pitman
Aquaforest

Aquaforest
Making Content Findable
Over 2000 Organizations in 50+ Countries rely on Aquaforest
Email: info@aquaforest.com Web: www.aquaforest.com
AQUAFOREST LTD - ENGLAND - COMPANY NO: 4344383
REGISTERED OFFICE: MIDSHIRES HOUSE, SUITE 32, MIDSHIRES BUSINESS PARK, LUTON, LU1 3AB

Asset Form PDF Days-4.pdf - Adobe Acrobat Pro

File Edit View Plug-Ins Window WindJack Tools Help

Open Create Print Signatures Comments

Tools Fill & Sign Comment

56.2%

New Asset Registration Form

2020 Edition Reference AB78

ID	Hal 9000	Category	Hardware
Condition	As New	Location	Berlin
Purchase Price	251363.00	Current Value	201250.00
Owner	Aquaforest	Acquired Date	1/1/2020
Approved By	J. K. Rowlinson		

Apps Aquaforest Limited... Aquaforest Limited... Searchlight - Docu... salesforce.com - Cu... https://www.aquafo...

Excel Neil Pitman > PDF Days Table Tools PDF Days Asset Register

File Home Insert Formulas Data Review View Help Design Tell me what you want to do

Undo Paste Cut Copy Format Painter

Clipboard Font Alignment Number

A6 Hal 9000

	A	B	C	D	E	F	G
1	Asset List						
2							
3	Description	Category	Condition	Location	Purchase Price	Current Value	Owner
4	Paper Shredder	Hardware	As New	Berlin	634	500	Aquaforest
5	75" Monitor	Hardware	As New	Berlin	634	500	Aquaforest
6	Hal 9000	Hardware	As New	Berlin	251363	201250	Aquaforest
7							
8							
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							
19							
20							
21							
22							

PDF Data for RPA : The Challenge



- As part of the RPA pipeline
- Extract, Process and Recognize PDF
- Structured and Semi-Structured data

PDF Data for RPA : Approaches

- Rely on Text Extraction Alone
- XMP
- Embedded
 - XML
 - JSON
- Tagged PDF
- Custom Metadata

PDF Text Extraction Issues

- Extractable Data
- Security
- Full Text
- Character Encoding
- Text Sequencing

```
<?xpacket begin="?" id="W5M0MpCehiHzreSzNTczkc9d"?>
<x:xmpmeta xmlns:x="adobe:ns:meta/" x:xmptk="Adobe XMP Core 5.4-c002 1.000000, 0000/00/00-00:00:00"
  <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    <rdf:Description rdf:about=""
      xmlns:xmp="http://ns.adobe.com/xap/1.0/"
      <xmp:CreatorTool>Picasa</xmp:CreatorTool>
    </rdf:Description>
    <rdf:Description rdf:about=""
      xmlns:mwg-rs="http://www.metadataworkinggroup.com/schemas/regions/"
      xmlns:stDim="http://ns.adobe.com/xap/1.0/sType/Dimensions#"
      xmlns:stArea="http://ns.adobe.com/xmp/sType/Area#"
      <mwg-rs:Regions rdf:parseType="Resource">
        <mwg-rs:AppliedToDimensions rdf:parseType="Resource">
          <stDim:w>912</stDim:w>
          <stDim:h>687</stDim:h>
          <stDim:unit>pixel</stDim:unit>
        </mwg-rs:AppliedToDimensions>
        <mwg-rs:RegionList>
          <rdf:Bag>
            <rdf:li rdf:parseType="Resource">
              <mwg-rs:Type></mwg-rs:Type>
              <mwg-rs:Area rdf:parseType="Resource">
                <stArea:x>0.680921052631579</stArea:x>
                <stArea:y>0.3537117903930131</stArea:y>
                <stArea:h>0.4264919941775837</stArea:h>
                <stArea:w>0.32127192982456143</stArea:w>
                <stArea:unit>normalized</stArea:unit>
              </mwg-rs:Area>
            </rdf:li>
          </rdf:Bag>
        </mwg-rs:RegionList>
      </mwg-rs:Regions>
    </rdf:Description>
    <rdf:Description rdf:about=""
      xmlns:exif="http://ns.adobe.com/exif/1.0/"
      <exif:PixelXDimension>912</exif:PixelXDimension>
      <exif:PixelYDimension>687</exif:PixelYDimension>
    </rdf:Description>
  </x:xmpmeta>
</xpacket>
```



OctoberPDFest ONLINE

Future for PDF in RPA

- PDF is a natural part of the RPA pipeline
- Aim for reliable “Machine Readable” PDF
- Use extraction technologies today if needed



See also
<https://www.pdfa.org/pdfcontainer-a-proposal/>

Session Summary

- Robotic Process Automation (RPA)
 - Uses software robots to mimic human actions and automate everyday tasks.

- This session covered
 - The place that PDF plays in RPA
 - How RPA fits in the wider spectrum of automation.
 - Typical use cases and demos
 - PDF-centric issues that need to be considered for successful implementations.

Aquaforest

Company and Products Overview

NEIL PITMAN

MANAGING DIRECTOR, AQUAFOREST LIMITED

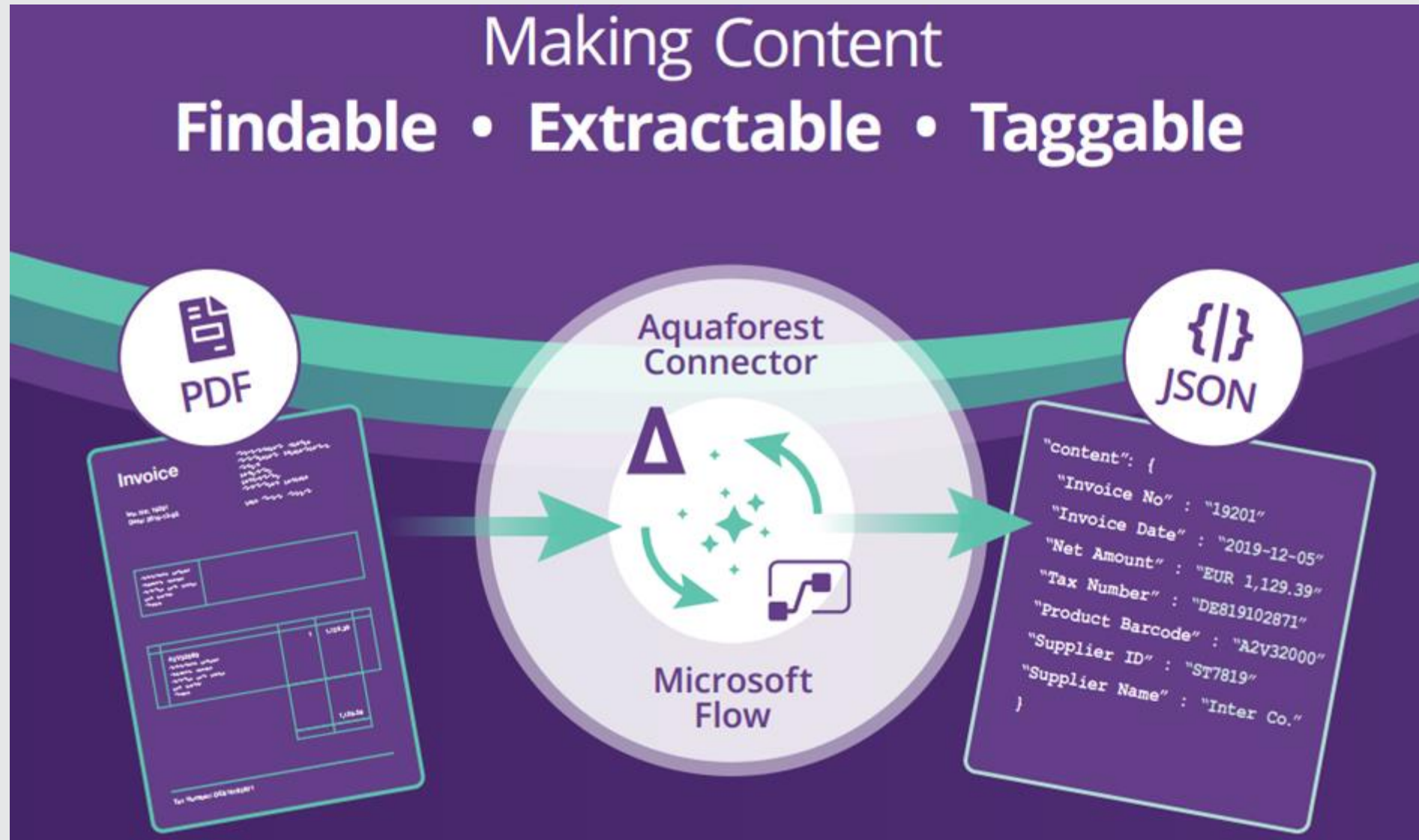
2,000 customers 50 countries



Flagship Products



Aquaforest PDF Connector



Aquaforest PDF Connector



Aquaforest Data Extractor

Load PDF File

n PDF Days-4.pdf

Chunks | Key-Value

Hide 'discarded results'

Office 365 (002).pdf

Chunks | Key-Value

Hide 'discarded results'

Microsoft

Invoice

October 2019

Invoice Date: 14/10/2019

Invoice Number: E0200996Q2

Due Date: 13/11/2019

3.00 GBP

Sold-To: Aquaforest Limited, Midshires House Suite 32, Smeaton Close, Aylesbury BUCKS HP19 8HL, United Kingdom, VAT: ip: 0029750476

Bill-To: Aquaforest Limited, Midshires House Suite 32, Smeaton Close, Aylesbury BUCKS HP19 8HL, United Kingdom

Service Usage Address: Aquaforest Limited, Midshires House Suite 32, Smeaton Close, Aylesbury BUCKS HP19 8HL, United Kingdom

Order Details

Product: Online Services

Customer PO Number: 04370403-37c6-4a28-97ee-f8dc528a63e8

Order Number: 14/09/2019 - 13/10/2019

Billing Period: Net 30

Payment Terms: 13/11/2019

Due Date: 13/11/2019

Payment Instructions: Please DO NOT PAY. You will be charged the amount due through your selected method of payment.

Billing Summary

Charges: 3.00

Discounts: 0.00

Credits: 0.00

VAT: Reverse Charge

Total: 3.00

Address

Bill-To: Aquaforest Limited, Midshires House Suite 32, Smeaton Close, Aylesbury BUCKS HP19 8HL, United Kingdom

Service Usage Address: Aquaforest Limited, Midshires House Suite 32, Smeaton Close, Aylesbury BUCKS HP19 8HL, United Kingdom

Postcode: HP19 8HL

Product: Number: Customer PO

Due Date: 13/11/2019

Invoice Number: E0200996Q2

Invoice Date: 14/10/2019

Aylesbury BUCKS HP19 8HL

When a file is created (Preview)

Get data from PDF

* File Content

File content x

Expected Keys

ID

Price

Page Range

A string representation of the page numbers you want to process. E.g 1,3-4

Show advanced options v

Run a UI flow for desktop (Preview)

* UI flow

UIF001

Edit

ID

1-ID x

Price

2-Price x

100



Aquaforest Searchlight


October**PDF**est ONLINE

Aquaforest
Searchlight

Making SharePoint
Content **Findable**

 Office 365

 SharePoint

 Microsoft
Azure



USE CASE Converting Archives to Fully Text Searchable PDF

Aquaforest Use Case Reference ADX/32-101

Aquaforest

INPUT FILES



Document Archive
Input File Headline



Server Automation



Autobahn DX Server is able to automate and manage large scale conversions with its auditable job manager. Support is provided for conversion from formats including TIFF, PDF, Word, Excel, MSG. PDF files are automatically analysed to determine whether OCR needs to be applied to make the file fully searchable.

Software & Systems:

Autobahn DX Server (1 – 32 Cores) • OS : Windows 10 or Server (2012 or later)
Platform : VM or Physical • RAM: 2GB per Core
Storage: 10GB + Temporary space at least equal to 2 x the size of the source document archive.
Performance guide : 1,000 pages per hour per Core

[Autobahn DX Homepage](#)



OUTPUT FILES



Fully Text Searchable PDF
Output File Headline

Aquaforest

Company and Products Overview

NEIL PITMAN

MANAGING DIRECTOR, AQUAFOREST LIMITED