

Separation Profile, Scan Profile, Extraction Groups, Classification Groups, and Folders

Work Like Tomorrow.™

KOFAX

Definitions: Structured, Semi-structured, Unstructured

- Kofax TotalAgility 7 includes modules and activities that work with Kofax Transformation Designer to provide advanced document classification, separation, and extraction for structured, semi-structured and unstructured documents

Known:

- ☒ Information
- ☒ Position

Structured
Forms

Known:

- ☒ Information
- ☒ Position

Semi-structured
Invoices/Orders

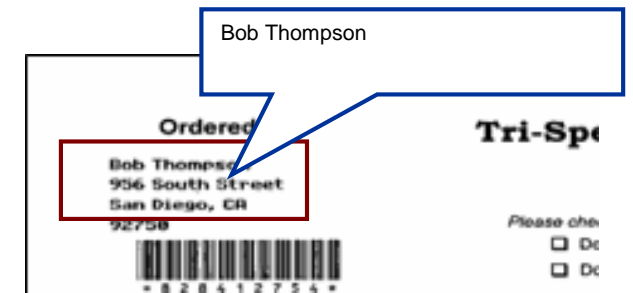
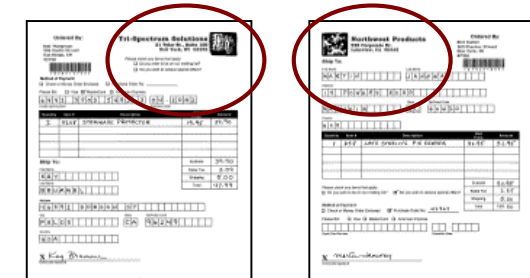
Known:

- ☒ Information
- ☒ Position

Unstructured
Correspondence

Classification, Separation, Extraction

- **Classification** – Distinguishing between different Document Types and Form Types.
- **Separation** – Creating boundaries between individual documents; documents may be single-page or multi-page.
- **Extraction** – The process of automatically lifting data from a document using rules or zones along with the output from a recognition engine (OCR, OMR, barcode, etc). This data usually is output to a database or document management application for later retrieval.



Page Alignment

Registration/anchors is the alignment or calibration of a page being imported or scanned.

Registration/anchors helps to assure that zones line-up with the goal of enhancing the accuracy of optical recognition operations.

Northwest Products
525 Corporate Dr.
Lakeview, CA 90435

Ordered By:
Bill Glaser
365 Planter Street
New York, NY
07326

Ship To:
First Name: MARTIN Last Name: JANEWAY
Address: 19 POWERS ROAD
City: MANHEIM State: IL Zip/Postal Code: 60420
Country: USA

Quantity	Item #	Description	Unit Price	Amount
1	038	LANG STERLING PIE SERVER	32.95	32.95

Please check any items that apply.
☐ Do you wish to be on our mailing list? ☒ Do you wish to receive special offers?

Method of Payment:
☐ Check or Money Order Enclosed ☒ Purchase Order No. 42367
Please Bill: ☐ Visa ☐ MasterCard ☐ American Express

Credit Card Number: Expiration Date:

X Martin Janeway
Authorized Signature

Northwest Products
525 Corporate Dr.
Lakeview, CA 90435

Ordered By:
Bill Glaser
365 Planter Street
New York, NY
07326

Ship To:
First Name: MARTIN Last Name: JANEWAY
Address: 19 POWERS ROAD
City: MANHEIM State: IL Zip/Postal Code: 60420
Country: USA

Quantity	Item #	Description	Unit Price	Amount
1	038	LANG STERLING PIE SERVER	32.95	32.95

Please check any items that apply.
☐ Do you wish to be on our mailing list? ☒ Do you wish to receive special offers?

Method of Payment:
☐ Check or Money Order Enclosed ☒ Purchase Order No. 42367
Please Bill: ☐ Visa ☐ MasterCard ☐ American Express

Credit Card Number: Expiration Date:

X Martin Janeway
Authorized Signature

Alignment

Page being imported or scanned



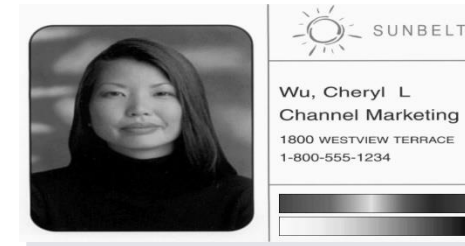
Sample Page

Image Concepts

- A color image
 - True color
 - 24 bits/pixel



- A grayscale image
 - 256 levels of gray
 - 8 bits/pixel



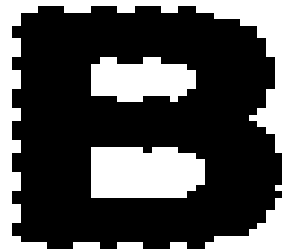
- A binary image
 - A black & white image
 - 1 bit/pixel
 - Very efficient file format



A pixel (picture element) is the smallest item of information in an image

Enhanced Images Lead to Improved Recognition

- Images can be enhanced through image clean-up inside Kofax TotalAgility
- Different recognition engines and image clean up options
- High-quality images result in less jagged edges and artifacts that produce crisper text
- Processing through Kofax image capture operations produce remarkably higher quality images resulting in significantly higher optical recognition accuracy



Results using traditional image capture technology



Results using Kofax Image Cleanup technology

Image Perfection using Virtual Rescan (VRS)

No VRS

USA Airbill 807030972286 0210 Form 10 No. Billing Copy

1 From _____ Sender's Account Number 4293888

Date _____

Sender's Name JANET ROSE Phone 562-726-0110

Company JANET'S FLOWERS

Address 285 LONG BEACH BLVD. Dept./Floor/Suite/Room

City LONG BEACH State CA ZIP 90802

2 Your Internal Billing Reference Information _____

3 To Dr Stevenson Phone _____

Company Expert Clinical Laboratory

Address 16245 Laguna Canyon Road Dept./Floor/Suite/Room

City Tustin State CA ZIP 92618

For HOLD at Location check here For WEEKEND Delivery check here

1 Hold Weekday 31 Hold Saturday 32 Hold Sunday 33 NEW Sunday Delivery

4a Express Package Service Packages under 150 lbs. Delivery commitment may be later in some areas

1 Priority Overnight 5 Standard Overnight

6 First Overnight 20 Express Saver

4b Express Freight Service Packages over 150 lbs. Delivery commitment may be later in some areas

7 Overnight Freight 8 2Day Freight 83 Express Saver Freight

5 Packaging 6 Letter 2 Pak 3 Box 4 Tube 1 Pig

6 Special Handling

7 Dry Ice

8 Release Signature

9 Payment

10 Total Packages Total Weight Total Declared Value Total Charges

VRS

AIRWAY USA Airbill 807030972286 0210 Form 10 No. Billing Copy

1 From _____ Sender's Account Number 4293888

Date _____

Sender's Name JANET ROSE Phone 562-726-0110

Company JANET'S FLOWERS

Address 285 LONG BEACH BLVD. Dept./Floor/Suite/Room

City LONG BEACH State CA ZIP 90802

2 Your Internal Billing Reference Information _____

3 To Dr Stevenson Phone 949-727-1735

Company Expert Clinical Laboratory

Address 16245 Laguna Canyon Road Dept./Floor/Suite/Room

City Tustin State CA ZIP 92618

For HOLD at Location check here For WEEKEND Delivery check here

1 Hold Weekday 31 Hold Saturday 32 Hold Sunday 33 NEW Sunday Delivery

4a Express Package Service Packages under 150 lbs. Delivery commitment may be later in some areas

1 Priority Overnight 5 Standard Overnight

6 First Overnight 20 Express Saver

4b Express Freight Service Packages over 150 lbs. Delivery commitment may be later in some areas

7 Overnight Freight 8 2Day Freight 83 Express Saver Freight

5 Packaging 6 Letter 2 Pak 3 Box 4 Tube 1 Pig

6 Special Handling

7 Dry Ice

8 Release Signature

9 Payment

10 Total Packages Total Weight Total Declared Value Total Charges

VRS

With VRS

AIRWAY USA Airbill 807030972286 0210 Form 10 No. Billing Copy

1 From _____ Sender's Account Number 4293888

Date _____

Sender's Name JANET ROSE Phone 562-726-0110

Company JANET'S FLOWERS

Address 285 LONG BEACH BLVD. Dept./Floor/Suite/Room

City LONG BEACH State CA ZIP 90802

2 Your Internal Billing Reference Information _____

3 To Dr Stevenson Phone 949-727-1735

Company Expert Clinical Laboratory

Address 16245 Laguna Canyon Road Dept./Floor/Suite/Room

City Tustin State CA ZIP 92618

For HOLD at Location check here For WEEKEND Delivery check here

1 Hold Weekday 31 Hold Saturday 32 Hold Sunday 33 NEW Sunday Delivery

4a Express Package Service Packages under 150 lbs. Delivery commitment may be later in some areas

1 Priority Overnight 5 Standard Overnight

6 First Overnight 20 Express Saver

4b Express Freight Service Packages over 150 lbs. Delivery commitment may be later in some areas

7 Overnight Freight 8 2Day Freight 83 Express Saver Freight

5 Packaging 6 Letter 2 Pak 3 Box 4 Tube 1 Pig

6 Special Handling

7 Dry Ice

8 Release Signature

9 Payment

10 Total Packages Total Weight Total Declared Value Total Charges

VRS

- Zonal or full page based
 - Image Cleanup Methods
 - Advanced Despeckle
 - Character Smoothing
 - Despeckle
 - Fill Line Breaks
 - Light Thicken Filter
 - Remove Lines
 - Smooth + Clean
 - Thicken Filter
 - Thinning Filter



Ordered By:
Bill Slater
365 Planter Street
New York, NY
107326

First Name

M	A	R	T	I	N					
---	---	---	---	---	---	--	--	--	--	--

Last Name * 6 7 3 4 2

J	A	N	E	W	A	Y				
---	---	---	---	---	---	---	--	--	--	--

1	9	P	O	W	E	R	S	R	O	A	D								
---	---	---	---	---	---	---	---	---	---	---	---	--	--	--	--	--	--	--	--

M	A	N	H	E	I	M					
---	---	---	---	---	---	---	--	--	--	--	--

I	L
---	---

6	0	4	2	0				
---	---	---	---	---	--	--	--	--

u	s	a								
---	---	---	--	--	--	--	--	--	--	--

☐ Do you wish to be on our mailing list? ☒ Do you wish to receive special offers?

☐ Check or Money Order Enclosed ☒ Purchase Order No. 42367

[illegible]

Expiration Date

12-18-98

Copyright © 1999 Kluwer Academic Publishers. All rights reserved.

OCR - Optical Character Recognition

- Converts printed characters from image bitmaps (pixels) into computer-readable text

- Segmentation (find text lines, then characters in lines)

HELLO WORLD

*OCR is not easy, but can
I find the letters?*

Let's see!!!

- Matching (analyze characters)

B B B B B B B B B B

O O O O O O O O O O

- Post-processing
 - Dictionaries, Tri-Gram Analysis
 - Context (neighbour characters)

Recognition & Extraction Capabilities

- Bar code recognition – Detecting and reading bar codes
- Optical Character Recognition (OCR) – Reading machine printed and hand printed characters
- Optical Mark Recognition (OMR) – Determining the status of checkboxes
- Intelligent Character Recognition (ICR) – for handwriting

The image shows a mail-order form from Northwest Products. Several areas are highlighted with red boxes to illustrate different recognition capabilities:

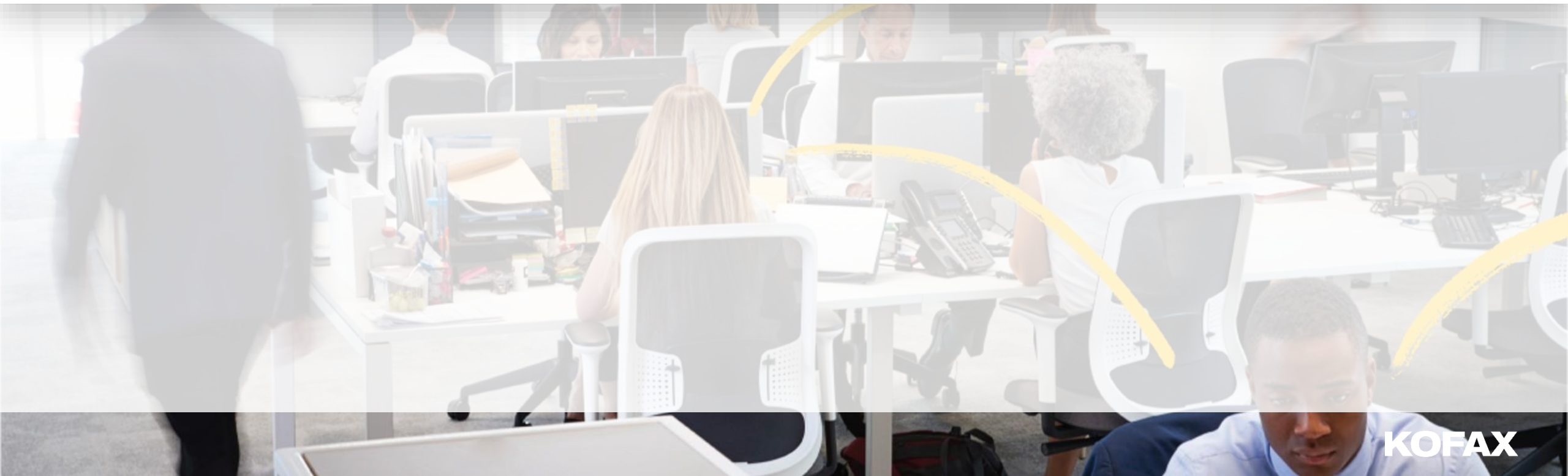
- A barcode at the top right.
- The company logo and address: Northwest Products, 525 Corporate Dr., Lakeview, CA 90435.
- The 'Ordered By:' field with the name Bill Slater, 365 Planter Street, New York, NY 87326.
- A barcode below the 'Ordered By:' field.
- The 'Ship To:' section, which includes:
 - First Name: MARTIN
 - Last Name: JANEWAY
 - Address: 19 POWERS ROAD
 - City: MANHEIM, State: IL, Zip/Postal Code: 60420
 - Country: USA
- Two checkboxes for mailing list and special offers, both of which are checked.
- The 'Method of Payment:' section, which includes:
 - Check or Money Order Enclosed (unchecked)
 - Purchase Order No. 42367 (checked)
 - Please Bill: Visa (unchecked), MasterCard (checked), American Express (unchecked)
 - Credit Card Number and Expiration Date fields.
- The 'Authorized Signature' field, which contains a handwritten signature: X Martin Janeway.

Extracted Data



- Can use Locators to locate and extract data from documents
- Extracted data can be validated and verified
- Extracted data can be passed to downstream processes (or exported to other back-end systems)

Separation Profiles



A Note on Folders and Separation

- Every time documents are imported or scanned they are automatically placed in a folder
- If you do not specify a design for the folder the Default folder will be used
- You can design your own folders which can contain the following:
 - Folder Name
 - Folder Fields
 - Subfolders
- If you do not specify folder separation, all documents will be placed in one top level folder when imported or scanned
- You can specify which folder is used to store the documents when you create a folder initialization variable in the process

Separation Profile

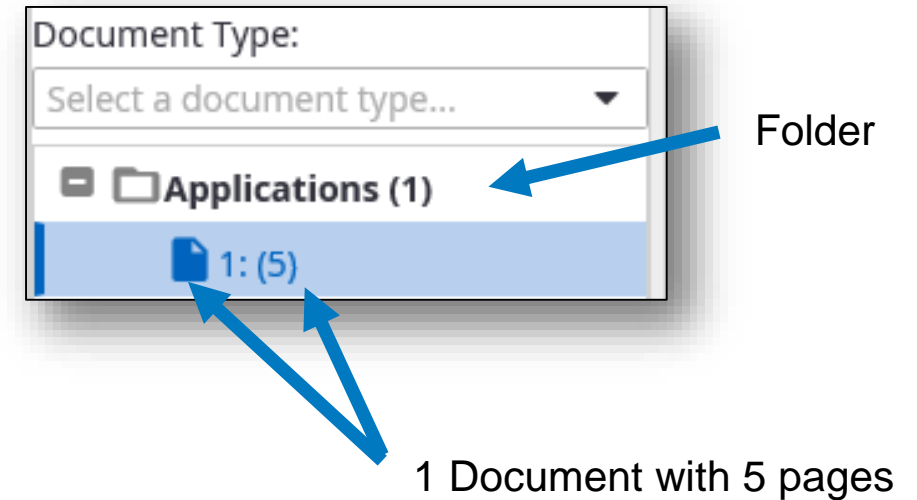
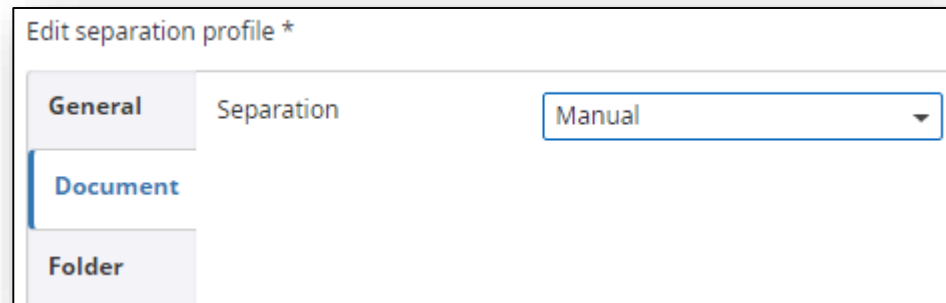
- A profile that tells the process (or the classification activity) how to separate documents
- You can specify document and folder separation
- Document Separation
 - A document is separated into single or multiple pages (Fixed value at Design time)
 - If a separation profile is applied at the process level, separation takes place on scanning
- Folder Separation
 - A group of documents is separated into a sub folder each time e.g. a barcode or patch code is found

Trainable Document Separation (TDS)

- Where separation profiles are not suited to your project, use:
- Trainable Document Separation (TDS)
 - For variable page numbers on multi-page documents, Separation can occur during Classification use Trainable Document Separation (TDS)

Example – Manual Separation

- A 5 page tiff file is imported
- Result on import is 1 Document with 5 pages



Example – Fixed Sheet (1 Sheet per Document)

- A 5 page tiff file is imported
- Result on import is 5 Documents with 1 page per document

Edit separation profile *

General	Separation	Fixed sheet
Document	Sheets per document	1
Folder		

Document Type:

Select a document type...

Applications (5)

- 1: (1)
- 2: (1)
- 3: (1)
- 4: (1)
- 5: (1)

Example – Fixed Sheet (2 Sheets per Document)

- A 5 page tiff file is imported
- Result on import is 3 Documents, 2 documents with 2 pages and 1 document with 1 page

Edit separation profile *

General	Separation	Fixed sheet
Document	Sheets per document	2
Folder		

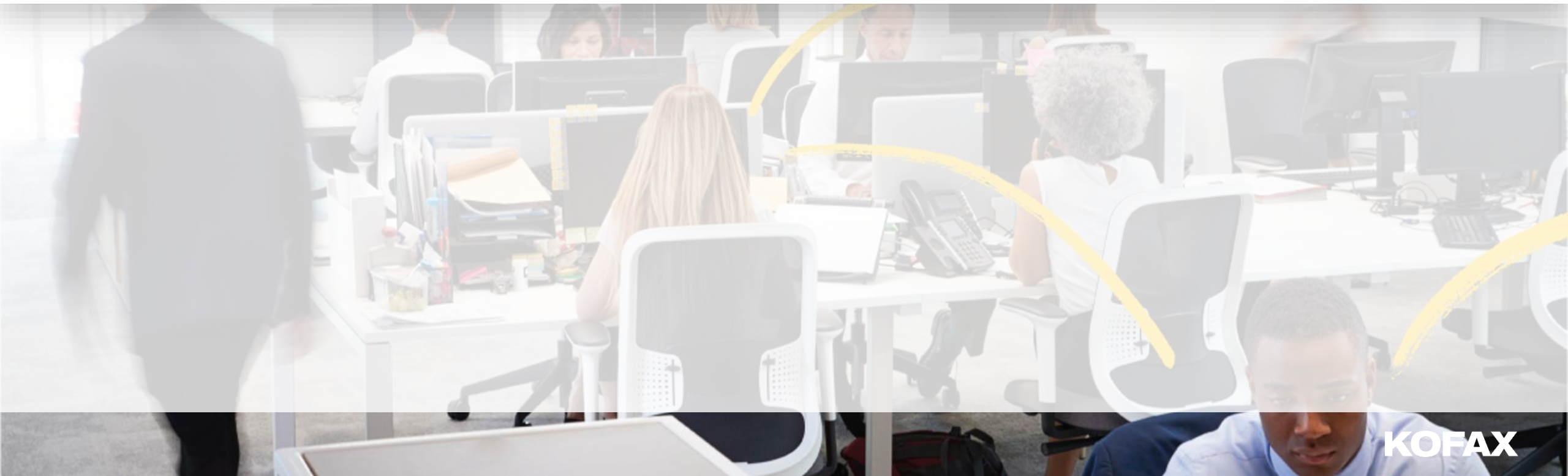
Document Type:

Select a document type...

Applications (3)

- 1: (2)
- 2: (2)
- 3: (1)

Scan/VRS Profile



Default Scan/VRS Profile

- A profile that can be applied to a process or Scan activity defining scan settings

Edit Scan / VRS profile

General * indicates a required field

Advanced

Name*

Description

Paper size

Orientation

Duplex

Paper source

Resolution

Color

Image enhancement settings ☐

Allow runtime editing

Paper size ☐

Duplex ☐

Resolution ☐

Color ☐

> **Process - capture applications**

i Classification group

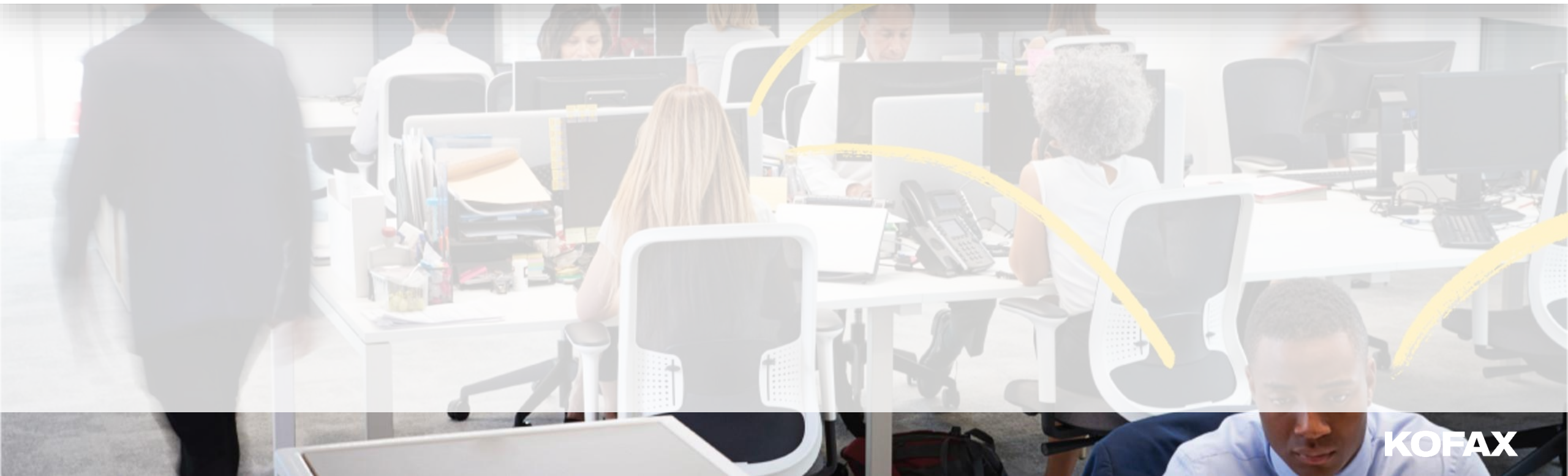
📄 Initialize from scan ☒

Separation profile

🔄 Scan/VRS profiles [Add](#) ↓ ↑

Default ×

Classification and Extraction Groups



What does Classification mean?

- We have three different document types
- Lets try to get the right group for these pictures:

Group A

???

???

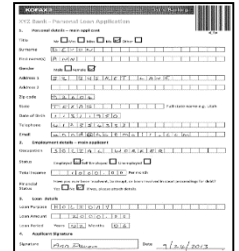
???

Group B

Group C

What does Classification mean?

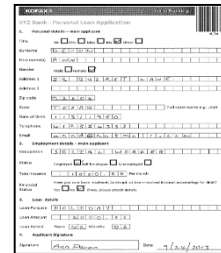
- We have three different document types
- Lets try to get the right group for these pictures:



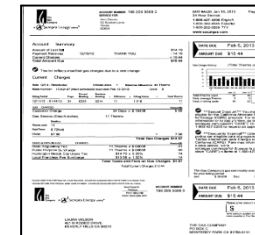
Group A



B



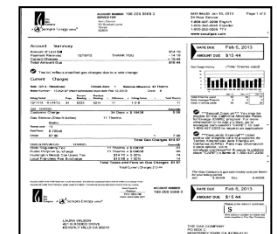
A



C



Group B



Group C

What is Classification?



- How could you match them?
 - Could you recognize forms?
 - Could you recognize a specific layout feature? (Fixed Layout)
 - Could you read some text (sentences)? (Semi Structured)
 - Could you read some words on it? (Unstructured)
- ➔ In our example Layout is fixed
- ➔ This is how layout classification works

Why do I need Classification?

- Classification allows the system to identify a document type so that during extraction the fields are extracted correctly
- Document Type: Application
 - Customer Name, Amount, Address, Zip Code, Employment Status
- Document Type: Drivers License
 - Expiry
 - Zip Code



KOFAX® Online Banking

XYZ Bank - Personal Loan Application

1. Personal details – main applicant

Title Mr ☐ Mrs ☐ Miss ☐ Ms ☒ Other ☐

Surname DEVON

First name(s) ANN

Gender Male ☐ Female ☒

Address 1 52 RUSSETT LANE

Address 2

Zip code 93604

State TEXAS Full state name e.g. Utah

Date of Birth 11/21/75

Telephone 412554338

Email anna@keltimaill.com

2. Employment details – main applicant

Occupation SCOTLAND WORKER

Status Employed ☒ Self Employed ☐ Unemployed ☐

Total Income 1000.00 Per month

Financial Status Have you ever been insolvent, bankrupt, or been involved in court proceedings for debt? Yes ☐ No ☒ If yes, please attach details.

3. Loan details

Loan Purpose HOLIDAY

Loan Amount 2000.00

Loan Period Years 2 Months 6

4. Applicant Signature

Signature Ann Devon Date 9/24/2013

Extraction Groups and Classification Groups

- A Classification group can consist of multiple extraction groups
- Extraction Groups define:
 - Document Types
 - Document Fields and Validation
- By adding extraction groups to a classification group we are telling the software to expect these document types when we classify a batch of documents from a capture source
- By splitting classification and extraction groups, we can foster re-use of document types

Extraction Groups and Classification Groups

Edit classification group

General * indicates a required field

Resources

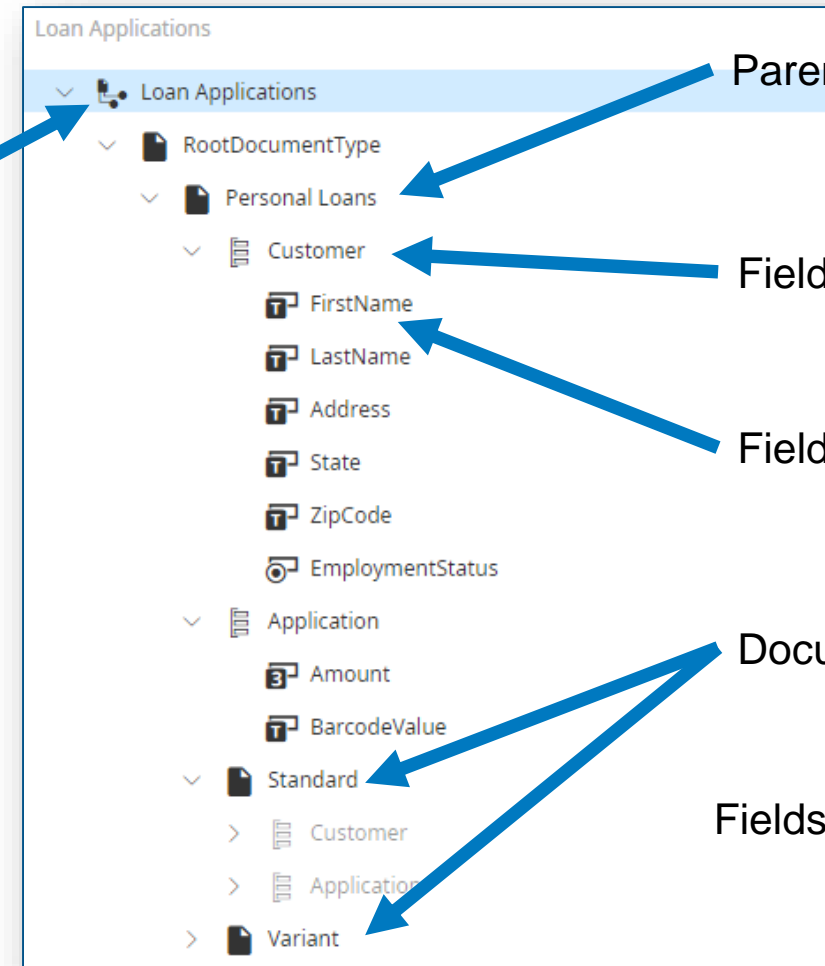
Name*

Description

Category*

Extraction groups*

[Add](#)



Parent Document Type

Field group

Fields

Document types

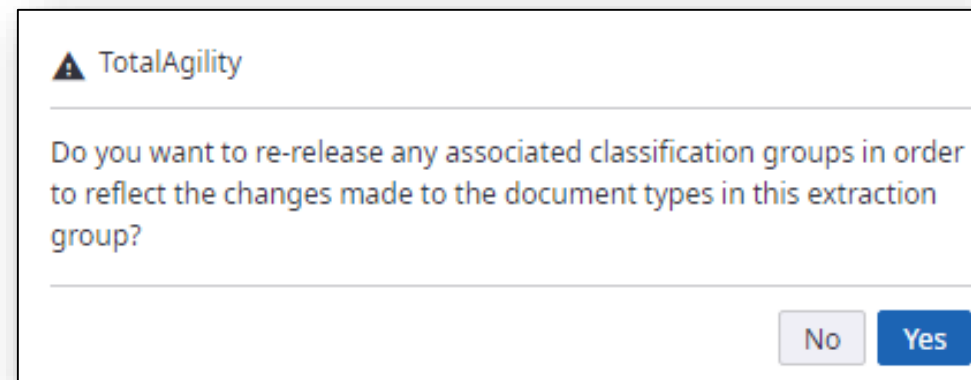
Fields inherited from Parent

Important Note

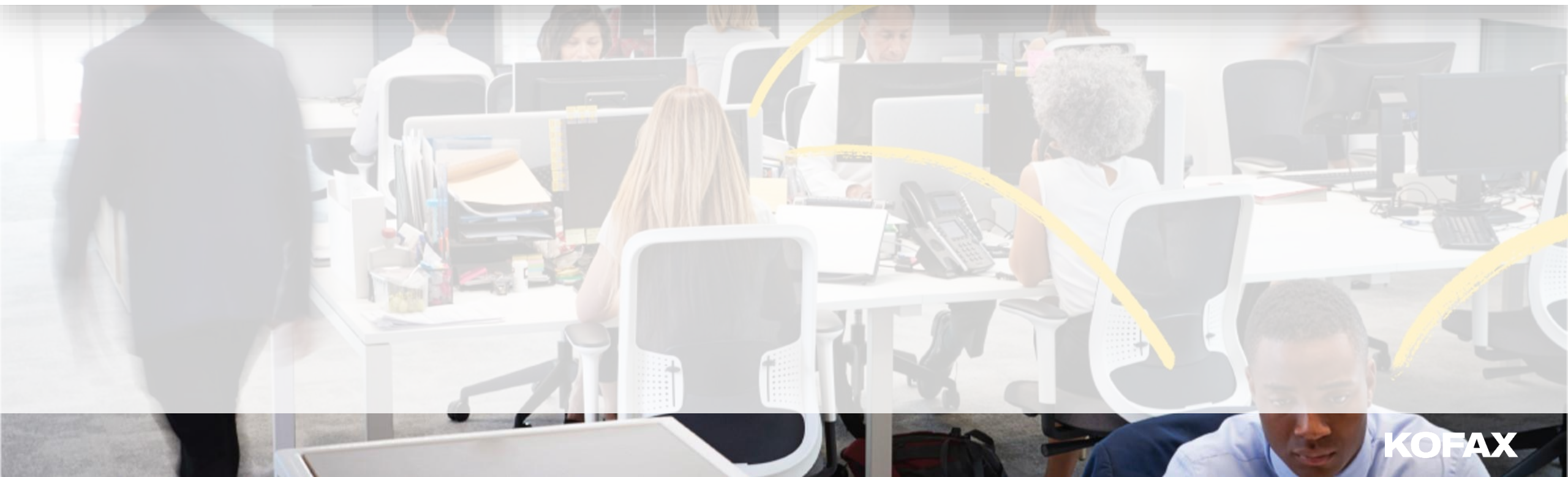
- If you make a change to an extraction group, you must re-release any classification groups that use that extraction group, if you wish the change to become effective



- You will be automatically prompted to update any classification groups when you release the extraction group



Folders



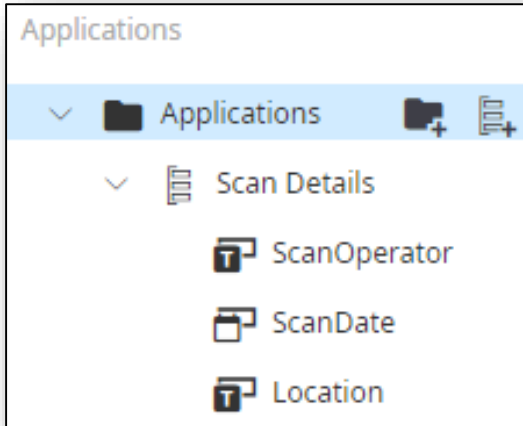
Folders



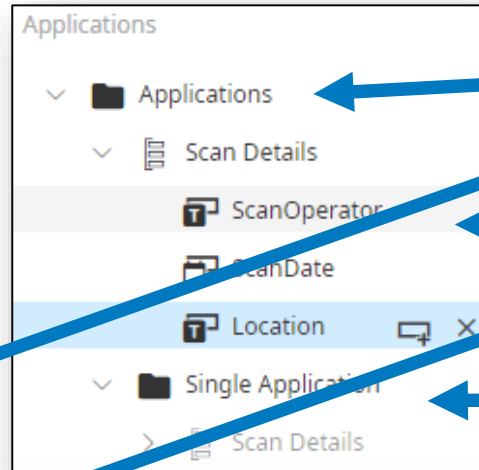
- When a batch of documents are scanned, the documents will be stored in a folder (or subfolders)
- You can design a folder by creating a new folder, allowing you to specify:
 - Folder Name and Folder fields e.g. Scan operator
 - Folder/Field validation e.g. a folder must have 3 documents
- You can choose to separate related documents into sub folders e.g. each time a bar code is found

Folders

Single Folder



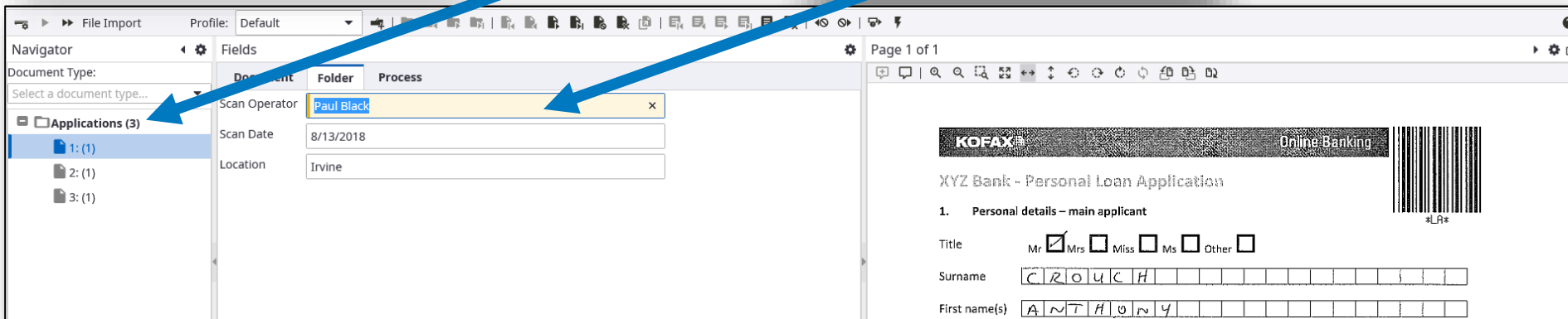
Folder with Subfolders



Top level folder

Folder fields

Sub Folder



Folder Variables

- You can determine which folder is used by a process (on scanning or import) by creating a folder initialization variable in the process

Applications Folder	APPLICATIONS_FOLDER	Folder	Applications	✓	⋮
-------------------------------------	---------------------	--------	--------------	---	---

- When a Scan create new job form is built for a specific process, the scan create new job form is connected to that process
- When you access the form and import or scan documents, the folder is created based on the design specified in the process initialization variable and references to the documents are placed in the folder