



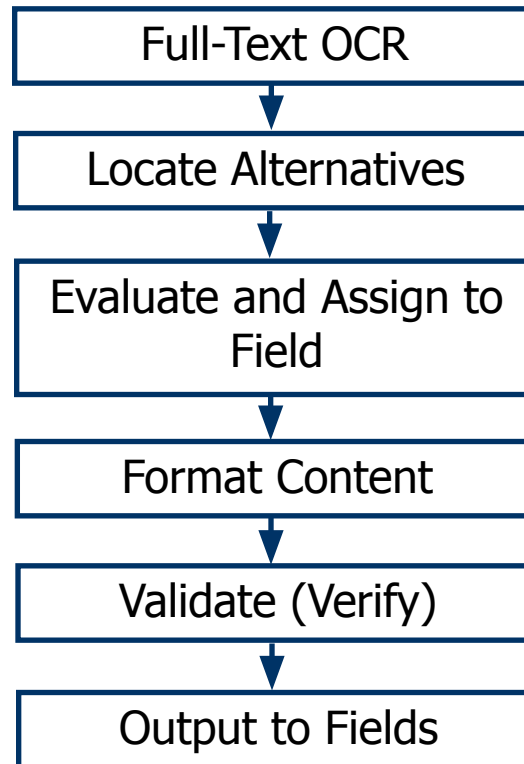
Настраиваемые (обучаемые) Локаторы для Счет-фактуры

Локаторы групп счетов, локаторы групп заказов и локаторы групп

Первый проект– Что дальше?

- ◆ Обзор KTM Extraction
- ◆ Обзор всех локаторов и анализаторов (evaluators)
- ◆ Настраиваемые (обучаемые) group locators (IGL, OGL, AGL, TGL)
- ◆ Анализатор (Evaluators) – используются для сравнения локаторов
- ◆ Некоторые более продвинутые локаторы
- ◆ Оптимизация форм
- ◆ Лучшие практики
- ◆ Написание сценариев
- ◆ Лицензирование

KTM's Extraction Process



Field	Content	Confidence
✓ VendorName	Oki Semiconductor	100.00 %
✓ VendorAddress1	785 NORTH MARY AVENUE	100.00 %
✓ VendorAddress2		100.00 %
✓ VendorCity	SUNNYVALE	100.00 %
✓ VendorState	CA	100.00 %
✓ VendorZip	94085-2909	100.00 %
✓ InvoiceNumber	310264 RI	95.00 %
✓ InvoiceDate	05/21/2003	95.00 %
✓ OrderNumber	15800051	100.00 %
✓ OrderDate		100.00 %
✓ DueDate		100.00 %
✓ Terms	NET THIRTY DAYS	95.00 %
✓ SubTotal	21553.42	90.00 %
✓ TaxAmount	0.00	100.00 %
✓ DateShipped		0.00 %
✓ ShippingHandling	0.00	100.00 %
✓ Total	21553.42	90.00 %
? LineItems		100.00 %

Примечание. Классификация (требуется для извлечения) происходит до или после полнотекстового OCR в зависимости от метода классификации.

Локаторы и Анализаторы

Basic	Advanced	Trainable	Legacy	Evaluators
Bar Code Locator	Database Locator	Amount Group Locator	Invoice Header Locator	Standard Evaluator
Advanced Zone Locator	Vendor Locator	Invoice Group Locator		Relation Evaluator
Format Locator	Table Locator	Order Group Locator		OCR Voting Evaluator
	Line Item Matching Locator	Trainable Group Locator		Database Evaluator
	Classification Locator	Text content locator		Address Evaluator
	A2iA Zone Locator			Advanced evaluator (formerly Invoice Evaluator)

Locators & evaluators in **bold black text** are covered in this level 1 course.

Обучаемые (настраиваемые) локаторы

- ◆ Amount Group Locator – Содержит поля которые относятся к сумме налога, общая сумма, и т. п. Обратите внимание что многие из этих полей необязательны и не должны присутствовать в счет-фактуре.
- ◆ Invoice Group Locator – Поиск информации по заголовку в счет-фактуре такой как номер счет-фактры, дата, имя и идентификатор поставщика
- ◆ Order Group Locator – Поиск информации, связанной с заказом, как номер заказа и дата заказа.
- ◆ Trainable Group Locator – Поиск информации в зависимости от настройки (обучения): общие, конкретные или оба, и не ограничивается счетами, но может использоваться практически для любого вида формы.
- ◆ Text Content Locator – на основе окружающего контекста. Полезно для неструктурированного документа, чтобы найти данные, которые вы не можете вернуть другим способом.
- ◆ Table Locator – Используется специальное (layout) обучение для возврата подробной информации о позиции из сложных счетов-фактур, которые не извлекаются должным образом в автоматическом режиме.

Другие локаторы и анализаторы (по алфавиту)

- ◆ Address Evaluator – сравнивает поля адресов с соответствующей базой данных и, если возможно, корректирует данные полей.
- ◆ Advanced Evaluator – принимает входные данные до трех локаторов в поле вывода в оценочные условия или «шаги» и возвращает значение. Может быть настроен для вывода на несколько полей.
- ◆ Advanced Zone Locator – считывает содержимое predetermined зон на фиксированных формах.
- ◆ Bar Code Locator – поиск и чтение штрих-кодов в документе.
- ◆ Classification Locator – Позволяет другим Kofax Transformation Modules project (с другой схемой классификации из текущего проекта) классифицировать документ и выводить результаты в поле. Например, можно определить проект, который классифицирует документы для 50 разных языков. Используя этот языковой проект, текущий документ может быть дополнительно классифицирован для определения языка в поле, которое назначено локатору.

Другие локаторы и анализаторы (по алфавиту)

- ◆ Database Locator – позволяет сопоставлять записи из данной базы данных с элементами документа. Должна использоваться плоская или «нечеткая» база данных со структурированными данными. Если база данных содержит данные клиента, локатор может идентифицировать имя, адрес и идентификатор клиента из документа, даже если документ может содержать даже не всю эту информацию.
- ◆ Database Evaluator – сравнивает результаты для полей, полученных из локатора зоны, в связанную базу данных.
- ◆ Format Locator – поиск элементов на основе регулярных выражений. Данные, которые обычно могут быть найдены с помощью этого типа локатора, включают суммы, даты и номера, такие как счет-фактура или страховой номер.
- ◆ Invoice Header Locator – принимает результаты от 4-х форматных локаторов, предоставляющих номера счетов, заказов, количества и даты и выдержки, формируя эти правильные значения для типичных данных заголовка счета, таких как номер счета, дата заказа, общие и налоговые значения.

Другие локаторы и анализаторы (по алфавиту)

- ◆ Line Item Matching Locator – сопоставляет позиции в счете-фактуре для позиций в ERP или другой базе данных SQL / ODBC.
- ◆ OCR Voting Evaluator – сравнивает результат зон с символом и выбирает лучший результат для каждого символа для сохранения в поле.
- ◆ Relation Evaluator – оценивает результаты одного локатора по сравнению с результатами другого локатора на основе относительного местоположения результатов.
- ◆ Script Locator – использует пользовательские события сценария WinWrap Basic для поиска данных. Локатор выходит на скрипт, который реализует метод определения местоположения или вызывает пользовательскую локализацию DLL.
- ◆ Standard Evaluator – сравнивает результаты нескольких локаторов и выбирает набор результатов на основе заданных критериев.
- ◆ Table Locator – заполняет поля таблицы. Доступны как ручные (основанные на шаблонах), так и автоматические методы извлечения на основе ключевых слов. Вы должны определить поля в «табличной модели», а затем сопоставить поля с локатором.
- ◆ Vendor Locator – Обнаруживает и оценивает данные, возвращаемые локатором базы данных, на основе дополнительной информации, такой как идентификатор поставщика, номер заказа на поставку, банковская информация итд

Обучаемые (настраиваемые) локаторы и база знаний

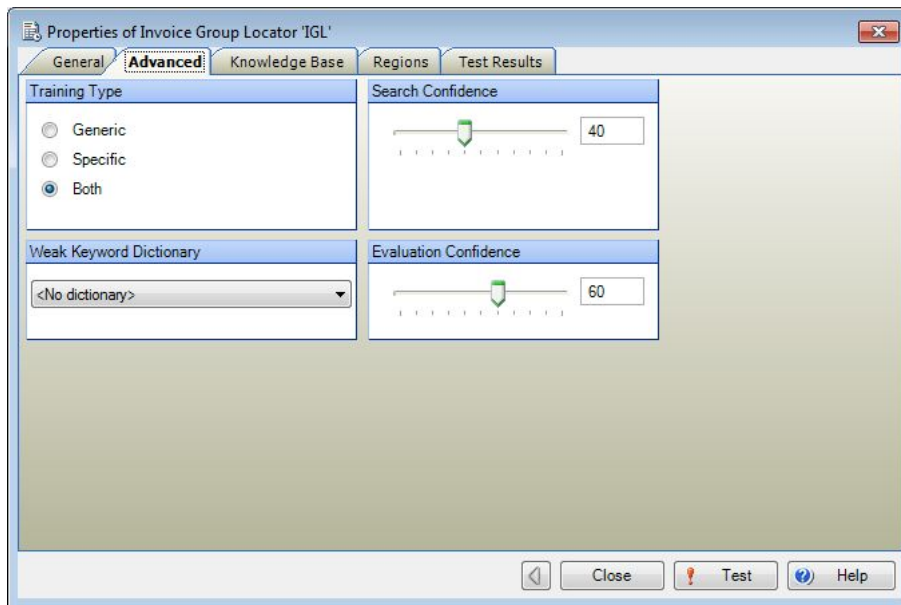
- ◆ IGL, OGL, AGL и TGL требуют обучения (так что локаторы текстового содержимого и некоторые локаторы таблиц). Обучение - это просто щелчок по слову или фразе на образце документов, чтобы заполнить поле, извлеченное обучаемым локатором.
- ◆ Когда вы подготовили достаточное количество образцов, вы можете создавать базы знаний из своего проекта.
- ◆ Базы знаний - это двоичные файлы специального назначения, которые заменяют ваши образцы учебных образцов и могут быть импортированы для использования другими проектами.
- ◆ Образцы обучения и базы знаний используют общие или конкретные алгоритмы.
- ◆ Общий алгоритм зависит от окружающих ключевых слов. По этой причине качество OCR важно. Он может использоваться в общем случае с помощью любого макета документа.
- ◆ Конкретный алгоритм зависит от компоновки конкретного документа.

Добавление счета в Группу Локаторов

Обратите внимание, что мы используем функцию DefaultDateFormatter для определения даты.

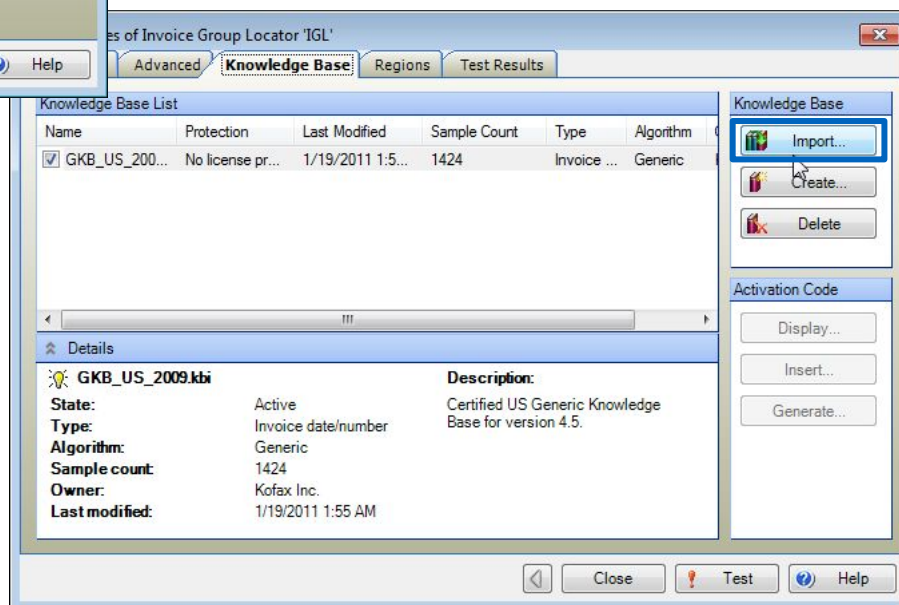
Classification Set	Filename	Classification Re	Confidence	Assigned Class
C:\...\ClassificationTraining	000001FC.xdc*	Oki	100.0%	
	000001FD.xdc*	Invoice		
	000001FE.xdc*	Invoice		Oki
	000001FF.xdc*			Oki
	00000200.xdc*	Invoices		Oki
	00000201.xdc*	Invoices		
	00000202.xdc			
	00000203.xdc			

Свойства групп локаторов



Тип обучения предназначен для использования как общего, так и специального обучения.
Примечание. Слабый словарь словаря и слайдеры проверки применяются только к общему обучению.

Мы будем импортировать один общий набор знаний, который мы предоставили вам для каждого из трех локаторов локаций, которые мы создадим. Поскольку они основаны на заранее подготовленных ключевых словах, это даст нам некоторые результаты прямо из коробки, без дополнительной подготовки. Но мы будем готовиться к лучшим

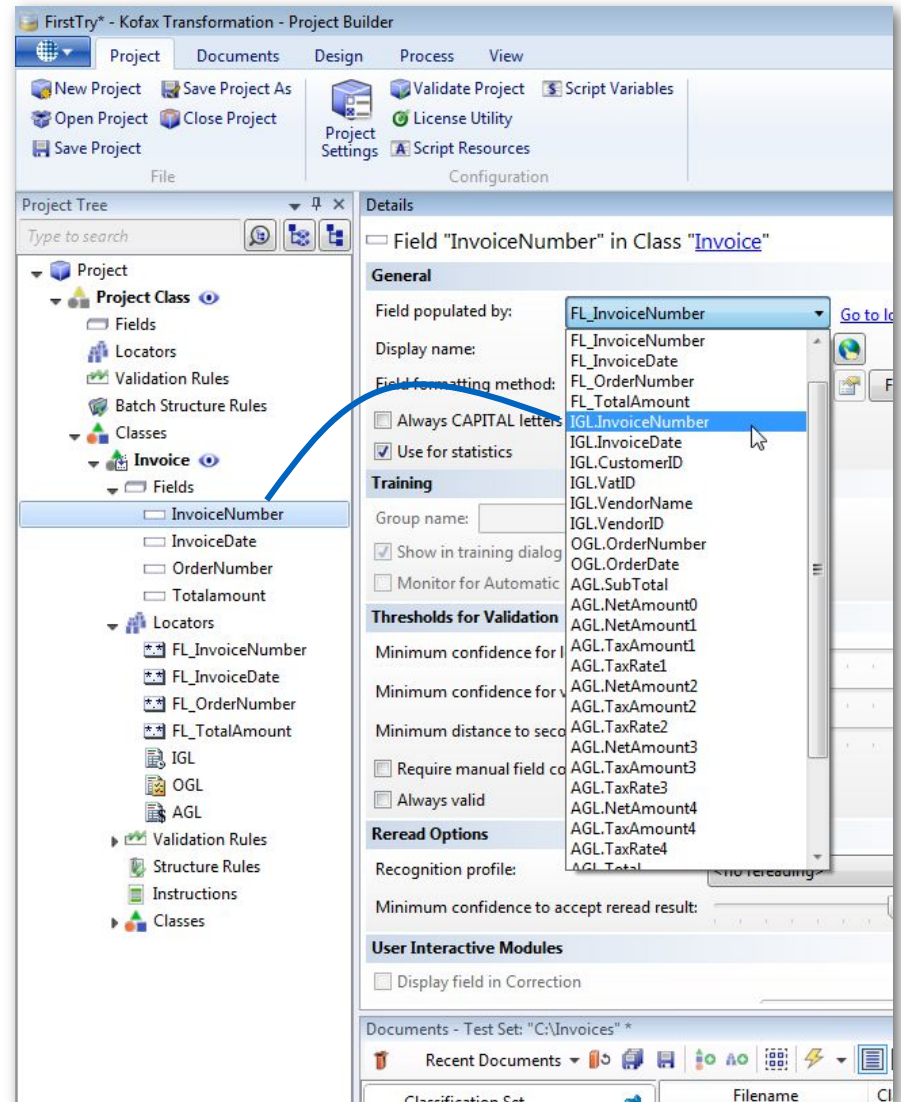
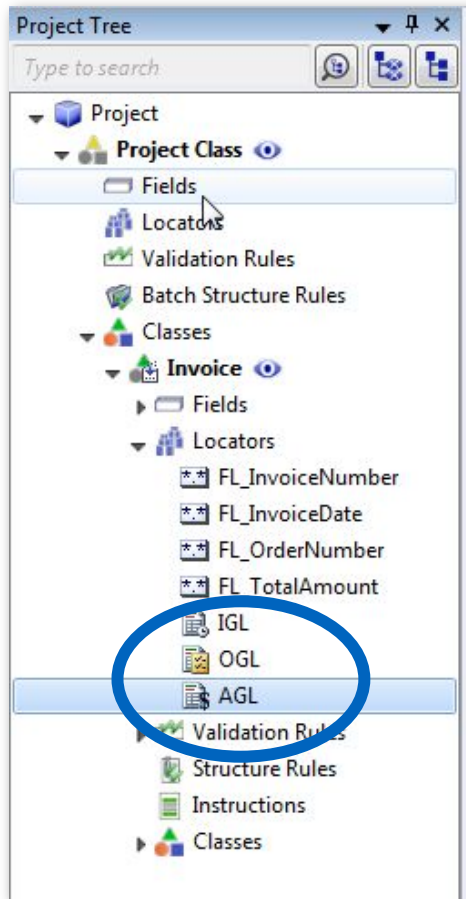


Добавлений локатора группы заказов и локатор групп суммы

Создайте еще два локатора, используя Locator Group Locator и методы Locator Group.

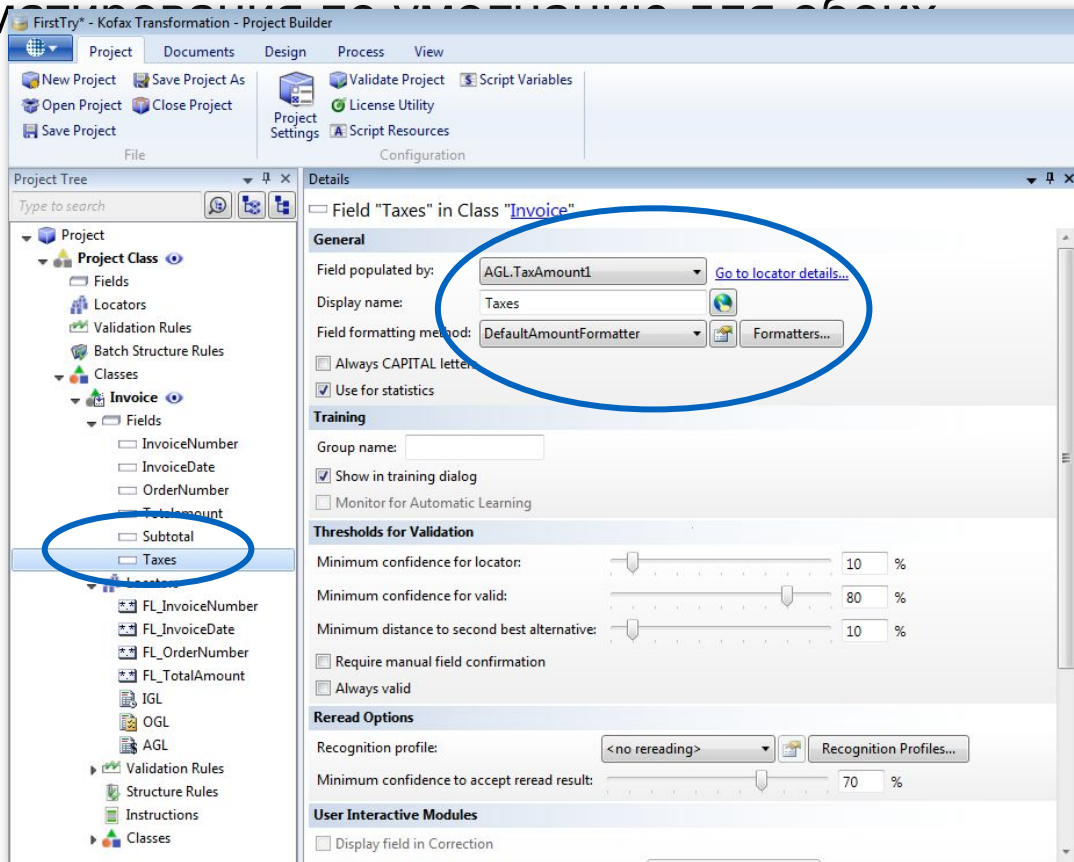
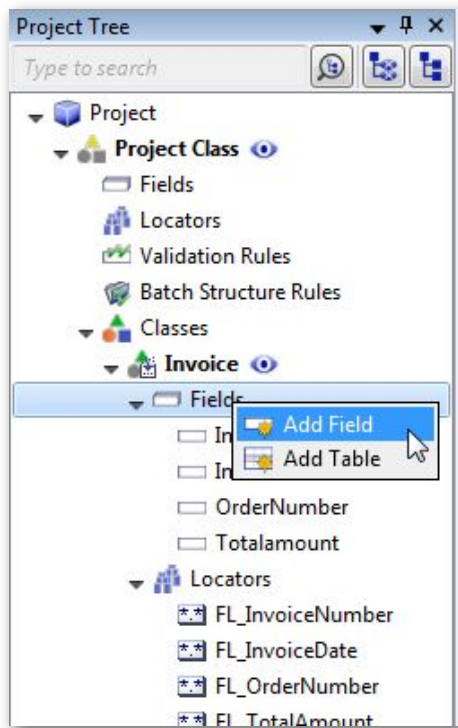
Затем давайте попробуем изменить вход локатора для четырех созданных нами полей. IGL вернет номер счета и дату, OGL вернет номер заказа. И AGL вернет общую сумму.

То, что мы пытаемся сделать, - использовать более «общий» метод для поиска наших данных в более широком разнообразии



Новые поля

- ◆ Пока мы это делаем, давайте выведем результаты для промежуточного итога и налогов. Это означает создание двух новых полей и вывод данных из наших локаторов. И мы применим формат форматирования для каждого из них.



Теперь нам нужно обучаться извлечению

Вы можете выбрать хорошие репрезентативные выборки из разных тренировок по извлечению макетов. Для конкретного обучения 1-4 выборки обычно достаточно. Для общего обучения важно получить образцы из как можно большего количества разных типов документов. Помните, что он основан на ключевых словах, и он должен знать все варианты, связанные с согласованными

The screenshot displays the Kofax software interface, divided into several panels:

- Details Panel (Left):** Shows configuration for a class named "Invoice" under "Project Class".
 - Classification:** Includes checkboxes for "Train this class for layout classification", "Train this class for content classification", "Valid classification result", and "Available for manual classification". A "Redirect to class" dropdown is set to "<no redirection>".
 - Subtree Classification:** Features an "Enable subtree classification" checkbox, a "Minimum confidence" slider at 40%, and a "Minimum distance" slider at 10%.
 - Trainable Document Separation:** Includes "Minimum page count" (1) and "Maximum page count" (0) fields.
 - Standard Document Separation:** Includes an "Ignore for separation" checkbox and radio buttons for "First page", "Middle page", and "Last page".
 - Recognition:** Includes a "Corresponding first page" dropdown set to "<none>".
- Document Viewer (Top Right):** Displays a scanned invoice from Oki Semiconductor. The classification result is "Invoices".
- Documents - Test Set (Bottom):** Lists files for training and testing, including "C:\...\ClassificationTraining", "C:\...\ExtractionTraining", and "C:\Invoices".
- Context Menu (Bottom Right):** A tooltip for the "Train for Extraction" icon explains: "Add the selected documents to the extraction training set for the selected class. This option is available only when a class is selected in the Project Tree and for all document sets except the extraction training set."

Пробуем в каждом поле левой кнопкой мыши или рисованием

Убедитесь, что курсор находится в правильном поле, а затем левой кнопкой мыши щелкните нужное значение для строк без пробелов или lasso значение, если пробелы включены.

Фиолетовыми значениями являются ключевые слова, используемые общим алгоритмом для

785 NORTH MARY AVENUE • SUNNYVALE, CA 94085-2909 • PHONE (408) 720-1900

INVOICE

INVOICE NO: 307243 RI
INVOICE DATE: 01/29/03
SHIPPED DATE: 01/29/03
CARRIER NAME: UPS GROUND
SO NUMBER: 556751 SO
PACKING SLIP NO.: 88345
00010

Okli Semiconductor

BILL TO: 195514
KOFAX IMAGE PRODUCTS, INC.
ACCOUNTS PAYABLE DEPT.
16245 LAGUNA CANYON ROAD
IRVINE CA 92618-3603

SHIP TO: 202180
KOFAX IMAGE PRODUCTS, INC.
16245 LAGUNA CANYON ROAD
IRVINE CA 92618-3603

CUSTOMER POI NO: 20020280
TERMS: NET THIRTY DAYS
F.O.B: FOB-SHIPPING PO
REF NO: 1031K
REP NAME: KOFAX IMAGE PRODUCTS
REFERENCE (SPSSL NO):

ITEM NO.	PART NUMBER	CUSTOMER PART NO.	U/M	QUANTITY	UNIT PRICE	EXTENSION
1.001	MSM99029-001LS	Tony	EA	1200	17.9500	21,540.00 USD

RECEIVED FEB 03 2003

ENTERED FEB - 6 2003

Please Remit To: OKI SEMI-SF LOCKBOX
P.O. BOX 7060
SAN FRANCISCO CA 94120
CUSTOMER

SUB TOTAL: 21,540.00 USD
SALES TAX: 3683* USD
AMOUNT DUE: 21,540.00 USD

*CUSTOMERS WHOSE PAYMENTS ARE NOT RECEIVED WITHIN 45 DAYS OF INVOICE DATE MAY BE SUBJECT TO CREDIT HOLD.

Начните с обучения на одном хорошо примере макета каждой формы.

Примечание. Чтобы исправить ошибочные ключевые слова, поместите курсор в соответствующее поле и [CTRL] щелкните правой кнопкой мыши по ключевому слову, чтобы очистить его, и, удерживая клавишу [CTRL], щелкните правой кнопкой мыши по правильному ключевому слову, чтобы установить его.

Добавить тренировочный комплект

The screenshot shows the 'Edit Document' window for a file named 'C:\Invoices\000001FE.xdc'. The menu bar includes 'File', 'Edit', 'View', and 'Help'. The toolbar contains 'Validate Document' and 'Add To Training Set'. The 'Fields' panel on the left lists several fields with checkboxes and values:

Field Name	Value
InvoiceNumber	307243 RI
InvoiceDate	01 / 29 / 03
OrderNumber	20020280
Totalamount	21,540.00
Subtotal	21,540.00'
Taxes	

The document preview on the right shows the header '785 NORTH MAI', the 'Oki Sem' logo, and the 'BILL TO:' information: 'KOFAX IMAGE E', 'ACCOUNTS PAYA', '16245 LAGUNA', and 'TRVTFE CA D'.

Примечание: На этой форме нет налогов.

Обучаемся на других документах

Edit Document - C:\Invoices\FedEx13.xdc

File Edit View Help

Validate Document Add To Training Set

Fields

Custom

InvoiceNumber 551 661 301

InvoiceDate 03 / 30 / 04

OrderNumber A5561I

Totalamount 243 72

Subtotal

Taxes

DUPLICATE 000030 **INVOICE** PAGE 01 OF 01

FedEx Freight (FXFE)

Send Payment to: FedEx Freight East P O Box 406708 Atlanta GA 30384-6708
 Direct Billing Inquiries to: P O Box 840, Harrison AR 72602-0840
 Phone: 1 870 741 9000 Fax: 1 870 395 4554 Toll-Free 1 866 393 4585

FREIGHT BILL NUMBER	
551 661 301	
TRACKING NUMBER	
DATE	
03/30/04	
ORIGIN	DEST
ATL	WHT
21264843	

SHIPPER 83963697 **BILL TO** 21264843
 DTM 4500 WICKERSHAM DR COLLEGE PARK GA 30337
 JOSLYN MANUFACTURING CO
 4 TRENDSET INC
 PO BOX 1208 MAULDIN SC 29662

CONSIGNEE 89267785	PO NUMBER	BL NUMBER	PAY
ABBEY CARPET OF ARCADIA DBA I J RAGER FLOOR COVER 52 E HUNTINGTON DR ARCADIA CA 91006	A5561I	010725	
ROUTING		REV	REV
VIKN		\$ 203.72	

PCS	HM	DESCRIPTION	WT(LBS)	N M F C	CLASS	RATE	TOTAL CHARGES
5		DISPLAY-CARPET/ CARPETING 8 47 PCF 1 SKIDS DISTRIBUTOR FOR BEAULIEU OF AMERICA 1 SMS STC 5 000360 FUEL SURCHG LTL SHPT1 507 550 LESS DISCOUNT DISCOUNT ARWF 40455 0000001 0006 *FXF 1000 06/30/03 LT 13607	309	057410-06	100	172.690	\$ 533.61
5		TERMS PPD	309				\$ 243.72

Reliable, Responsive Regional and Interregional LTL

Remittance Advice
 Please Return This Portion With Your Payment
 Terms Net 15 Days

PAYING PARTY 21264843
 #BNFJ2
 JOSLYN MANUFACTURING CO
 4 TRENDSET INC
 PO BOX 1208 MAULDIN SC 29662

FREIGHT BILL NUMBER
551 661 301
DATE
03/30/04
PLEASE PAY THIS AMOUNT
243 72

FedEx Freight (FXFE)

Send Payment to: FedEx Freight East
 P O Box 406708 Atlanta GA 30384-6708
THANK YOU !!

TRACKING NUMBER

Document has to be validated

300 DPI / 300 DPI

Добавляем другой

Edit Document - C:\Invoices\00000253.xdc

File Edit View Help

Validate Document Add To Training Set

Fields

Custom

Add To Training Set

InvoiceNumber 341350479


InvoiceDate 05 / 08 / 03

OrderNumber 20036106

Totalamount 1,053.80

Subtotal 978.00

Taxes 75.80



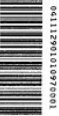
This is your INVOICE

1496* Page: 1 of 2

PID Number: 742618805
 Sales Rep: JOHN A COOK
 For Sales: (800)234-9999
 Sales Fax: (800)686-0438
 For Customer Service: (800)234-9999
 For Technical Support: (800)234-9999
 Dell Online: http://www.dell.com

Customer Number: 000318193
 Business Order: 20036106
 Order Number: 241350479
 Order Date: 05/01/03
 15 01 0 01 01 N

Invoice Date: 05/08/03
 Payment Terms: NET DUE 30 DAYS
 Shipped Via: 2ND DAY
 Waybill Number: 48190298662



SOLD TO:
 #60/NH/KPI
 2003 1515 200 00001097 1 AB 0.301 01
 ACCOUNTS PAYABLE
 KOFAX IMAGE PRODUCTS
 ACCOUNTS PAYABLE
 19245 LAGUNA CANYON RD
 IRVINE CA 92618-3603

SHIP TO:
 RECEIVING
 KOFAX IMAGE PRODUCTS
 19245 LAGUNA CANYON RD
 IRVINE, CA 92618-3601

PLEASE REVIEW IMPORTANT TERMS & CONDITIONS ON THE REVERSE SIDE OF THIS INVOICE


Order	Shipped	Item Number	Description	Unit	Unit Price	Amount
1	1	221-0985	Dimension 4550 Series, Intel Pentium 4 Processor at 2.53GHz	EA	868.00	868.00
1	1	311-2076	512MB DDR SDRAM at 333MHz	EA	0.00	0.00
1	1	310-1582	Dell Quiet Key Keyboard	EA	0.00	0.00
1	1	320-3000	Video ready option w/o monitor	EA	0.00	0.00
1	1	320-0452	32MB ATI RAGE ULTRA 4X AGP Video	EA	0.00	0.00
1	1	340-7886	60GB Ultra ATA/100 Hard Drive 7200RPM	EA	0.00	0.00
1	1	340-9626	3.5in Floppy Drive	EA	0.00	0.00
1	1	313-7222	Dell Application Back-up CD, Factory Install	EA	0.00	0.00
1	1	412-0306	Dell Support 2.0 for Dimension 4550	EA	0.00	0.00
1	1	420-1921	Microsoft Windows XP Home Edition, Service Pack 1, English	EA	0.00	0.00
1	1	310-1413	Logitech Optical USB Mouse	EA	0.00	0.00
1	1	430-0412	Intel Pro 120M Integrated PCI NIC Card	EA	0.00	0.00
1	1	313-3607	No modem requested for Dell Dimension	EA	0.00	0.00
1	1	313-1368	48x24x48x CD-RW Drive	EA	0.00	0.00
1	1	313-0947	Integrated ADI 1865 Audio	EA	0.00	0.00
1	1	313-4514	No Speaker Requested	EA	0.00	0.00
1	1	412-0326	NETWORK ASSOCIATES MCAFFEE.COM OEM ENGLISH 30 DAY TRIALEA	EA	0.00	0.00
1	1	481-8386	FACTORY INSTALL	EA	0.00	0.00
1	1	481-8386	No Digital Music Software requested	EA	0.00	0.00
1	1	481-8386	No Digital Imaging Software requested	EA	0.00	0.00
1	1	412-0380	Real Network RealOne Player, Basic, Version 6.0, English	EA	0.00	0.00
1	1	412-1367	No Productivity Software requested	EA	0.00	0.00
1	1	950-1390	*Type 3 Contract - Next Business Day Parts and Labor On-Site Response, Initial Year	EA	0.00	0.00
1	1	950-1392	*Type 3 Contract - Next Business Day Parts and Labor On-Site Response, 2YR Extended	EA	0.00	0.00

37399 * 886.00

01-1570-200 * 185.80

Ship. &/or Handling	\$	110.00
Subtotal	\$	978.00
Taxable	\$	978.00
Tax	\$	75.80
Invoice Total	\$	1,053.80

ENTERED MAY 14 2003



DETACH AT PERFORATION AND RETURN WITH PAYMENT

MAKE CHECK PAYABLE/REMIT TO:

DELL MARKETING L.P.
C/O DELL USA L.P.
DEPT. LA21205
PASADENA, CA 91185-1205

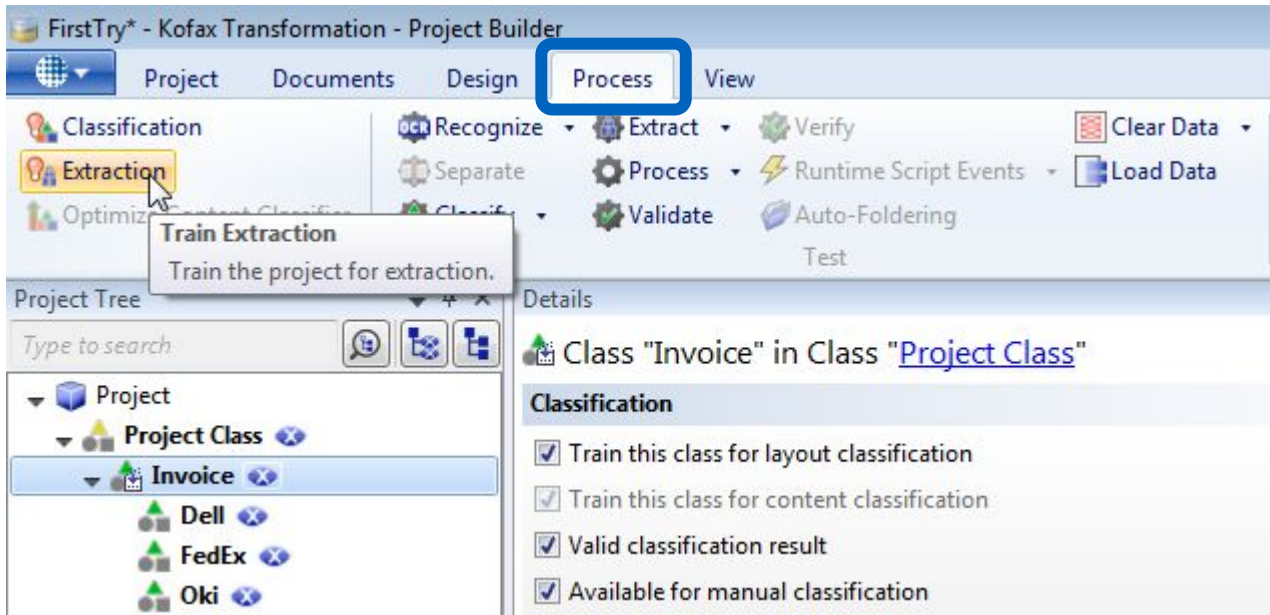
Invoice Number: 341350479
Customer Number: 000318193
Purchase Order: 20036106
Order Number: 341350479

(REL. Rev. 9/2002) 00034135047900000001053601500003161936

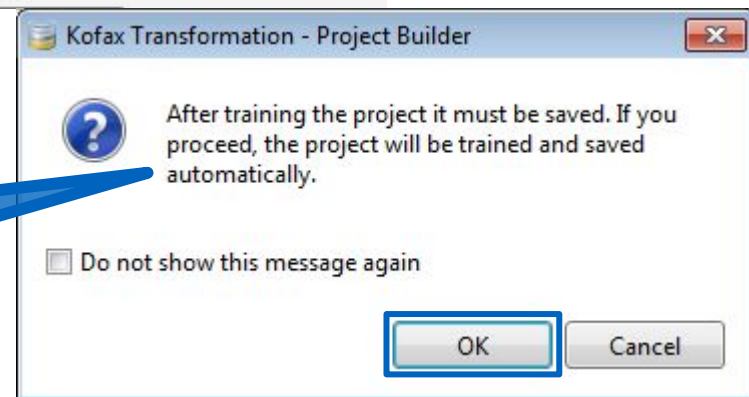
Document has to be validated

300 DPI / 300 DPI

Обучаемся на проекте



**Автосохранение.
Спасибо, Project
Builder!**



И тестируем

Project Tree

- Project
 - Project Class
 - Invoice
 - Dell
 - FedEx
 - Oki

Details

Class "Invoice" in Class "Project Class"

Classification

- Train this class for layout classification
- Train this class for content classification
- Valid classification result
- Available for manual classification

Redirect to class: <no redirection>

Subtree Classification

- Enable subtree classification

Minimum confidence: [slider]

Minimum distance: [slider]

Subtree classification via parent class required: <no restriction>

Subtree Classifier: [Properties...]

Extraction Results - Dell

Field	Content	Confidence
✓ InvoiceNumber	345004270	90.00 %
✓ InvoiceDate	05/08/2003	90.00 %
✓ OrderNumber	20036114	90.00 %
✓ Totalamount	1404.28	90.00 %
✓ Subtotal	1303.23	90.00 %
✓ Taxes	101.05	90.00 %

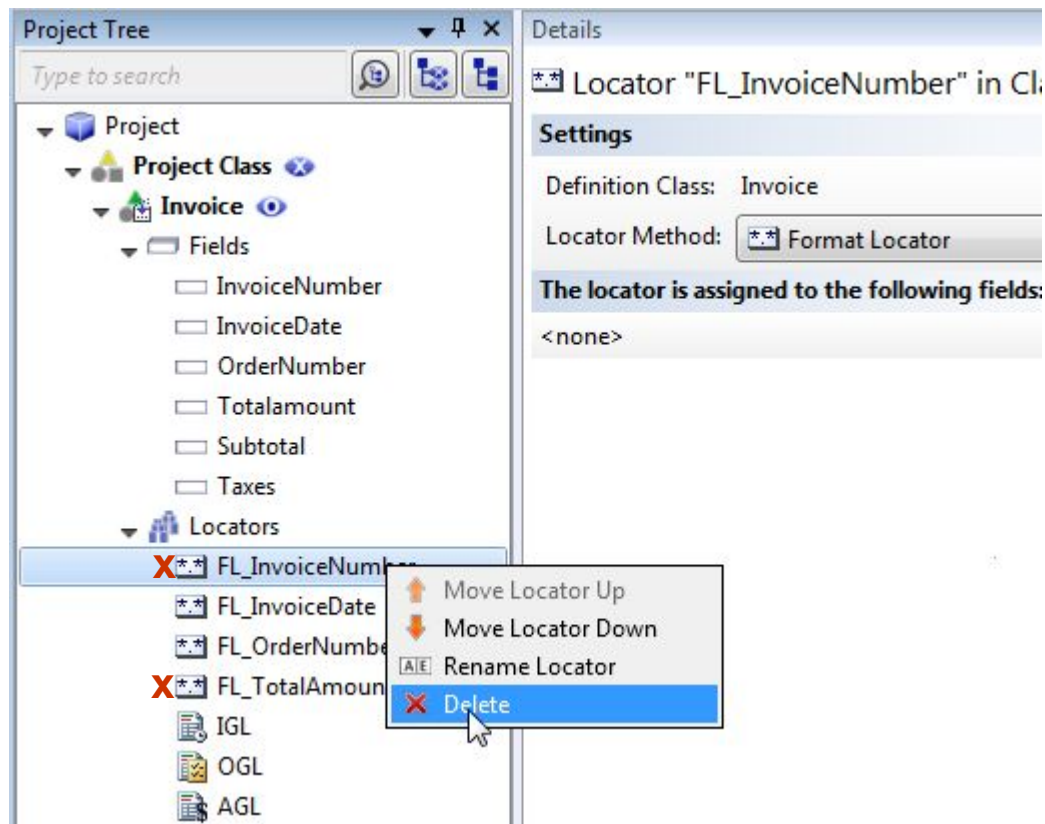
Documents - Test Set: "C:\Invoices"

Classification Set	Filename	Classification Re	Confidence	Assigned Class
C:\...\ClassificationTraining	0000024B.xdc			
	0000024C.xdc			
	0000024D.xdc			
	0000024E.xdc			
	0000024F.xdc			
	00000250.xdc			
	00000251.xdc*	Dell	85.9%	
	00000252.xdc			
	00000253.xdc*	Invoice		

Не забывайте, что вы можете протестировать локатор каждой группы отдельно, и вы можете (и должны) запустить Extraction Benchmark, чтобы проверить результаты вашего извлечения.

Удаляем неиспользуемые форматы локаторов

- ◆ Поскольку мы больше не используем локаторы формата, чтобы возвращать результаты, мы собираемся удалить пару из них. Мы оставим пару на месте для использования со стандартным оценщиком, о котором мы узнаем немного позже...



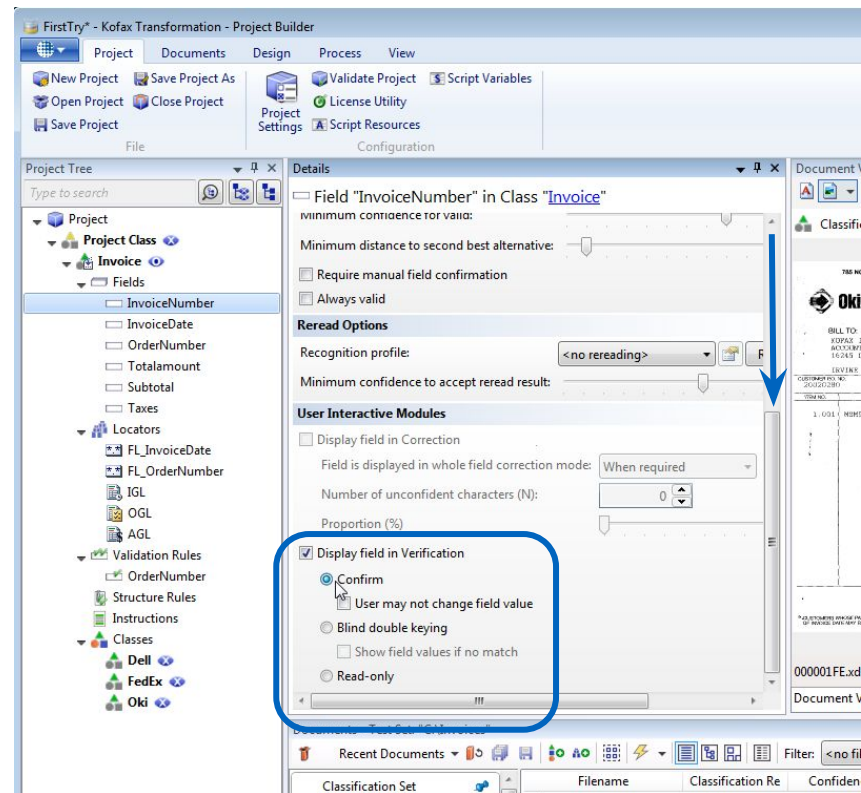
Добавляем КТМ Верификацию

- ◆ Модуль верификации позволяет верифицировать данные который уже были на валидации. Он является необязательным и должен использоваться только тогда, когда абсолютная точность некоторых полей является критичной. Верификация настроена для отдельных полей через Details Panel.

We'll turn on Verification for the InvoiceNumber, InvoiceDate and Totalamount fields.

- ◆ Существует три режима верификации:

- ◆ Подтверждение требует от оператора на валидации подтвердить путем нажатия [Enter].
- ◆ Blind double keying закрывает валидацию и требует от оператора ввода значения вручную. Затем сравниваются два значения.
- ◆ Только для чтения отображается подтвержденное значение для поля, но не позволяет оператору изменять его.



Демонстрация и задание