KOFAX

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

Full-Text OCR

↓

Locate Alternatives

↓

Evaluate and Assign to Field

↓

Format Content

↓

Validate (Verify)

↓

Output to Fields

Extraction Results - Oki

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.

KOFAX

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 предопределенных 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.

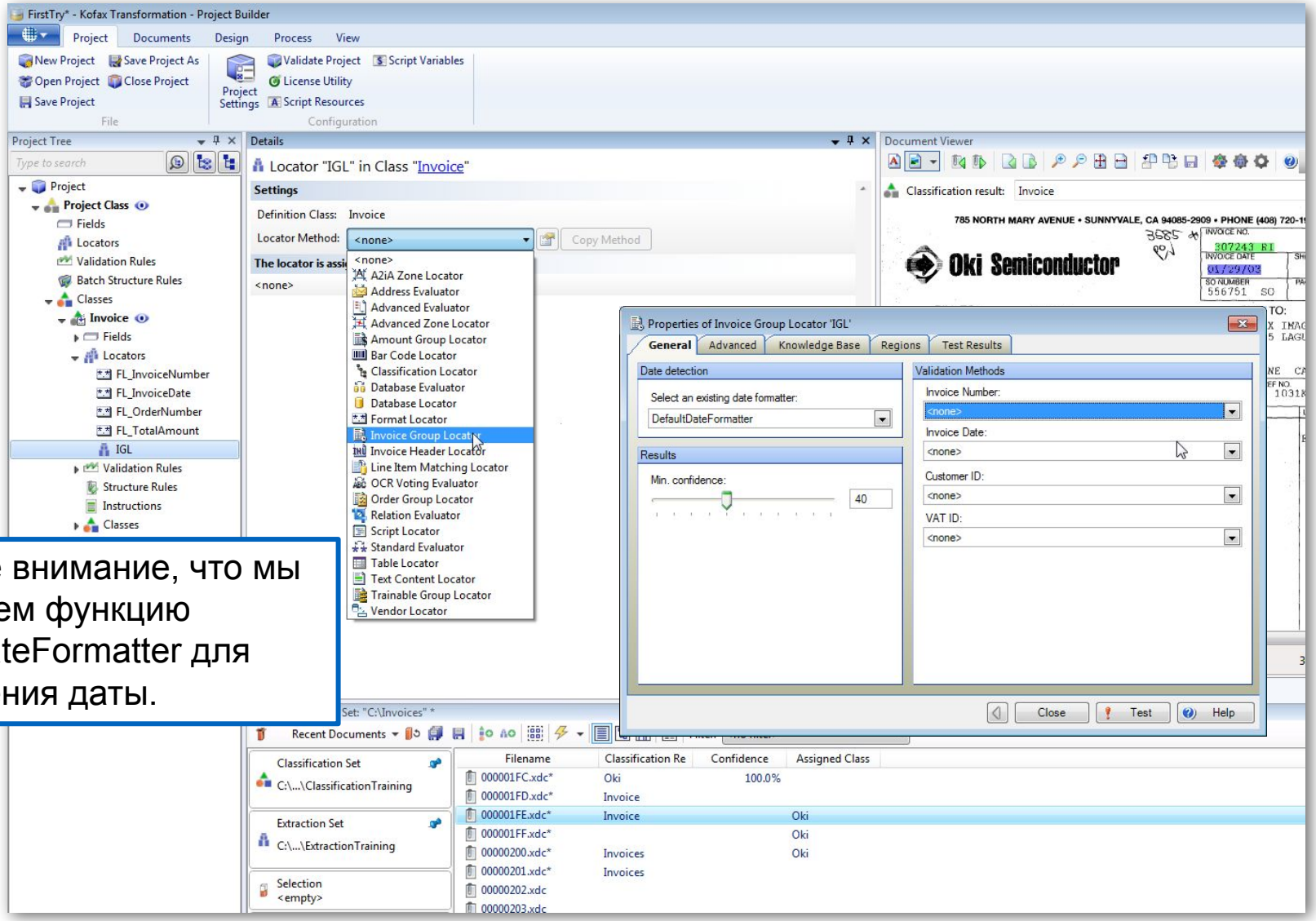
Другие локаторы и анализаторы (по алфавиту)

- 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

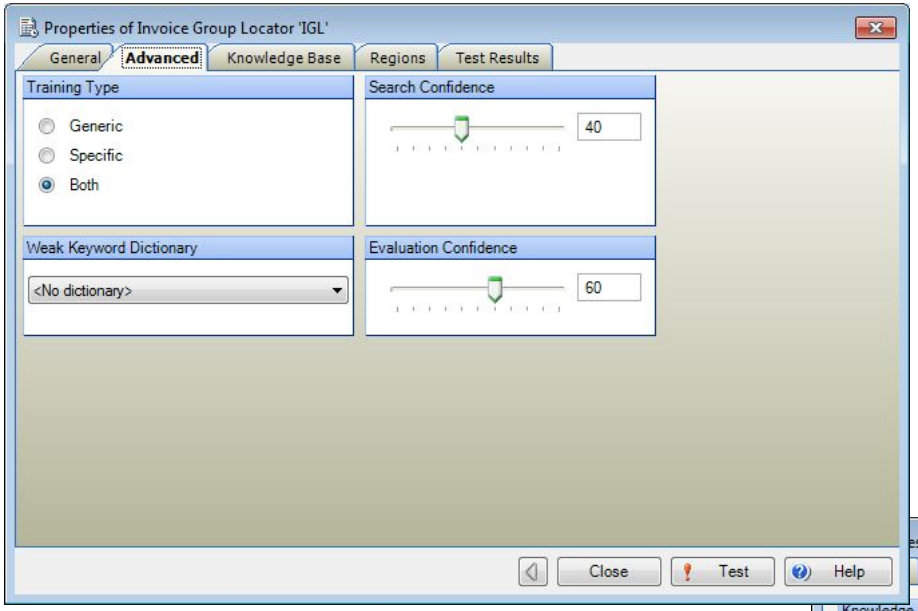
- IGL, OGL, AGL и TGL требуют обучения (так что локаторы текстового содержимого и некоторые локаторы таблиц). Обучение - это просто щелчок по слову или фразе на образце документов, чтобы заполнить поле, извлеченное обучаемым локатором.
- Когда вы подготовили достаточное количество образцов, вы можете 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

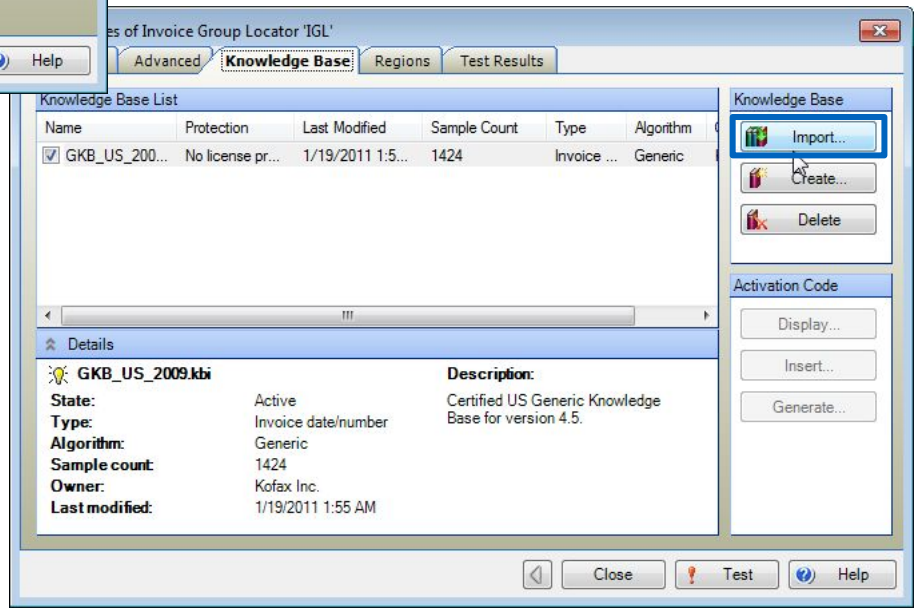


Обратите внимание, что мы используем функцию DefaultDateFormatter для определения даты.

Свойства групп локаторов



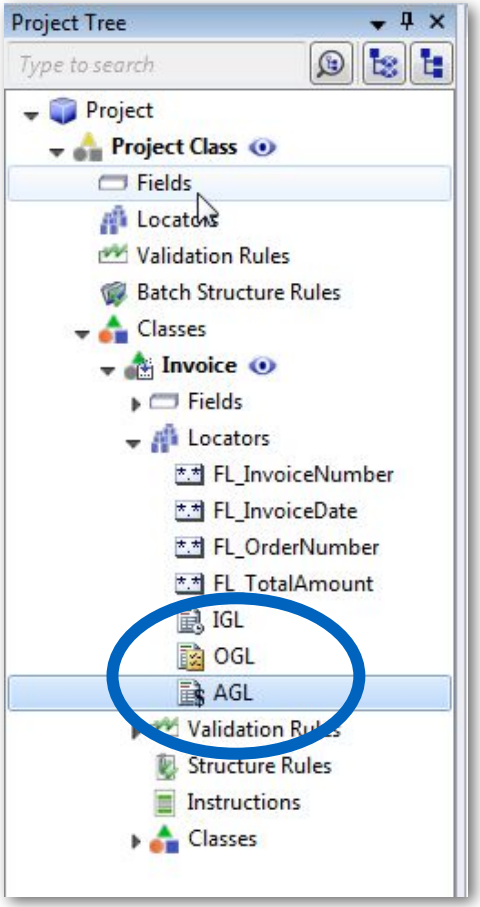
Тип обучения предназначен для использования как общего, так и специального обучения. Примечание. Слабый словарь словаря и слайдеры проверки применяются только к общему обучению.



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

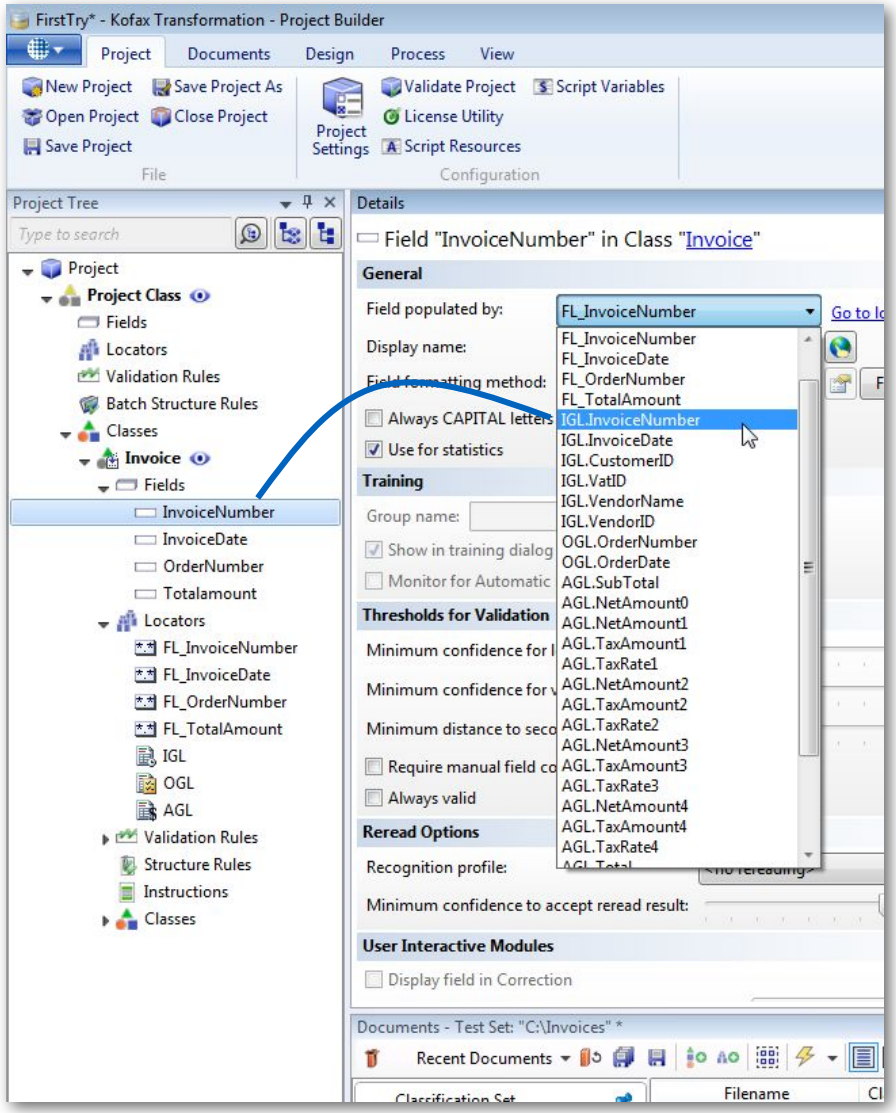
Добавлений локатора группы заказов и локатор групп суммы

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



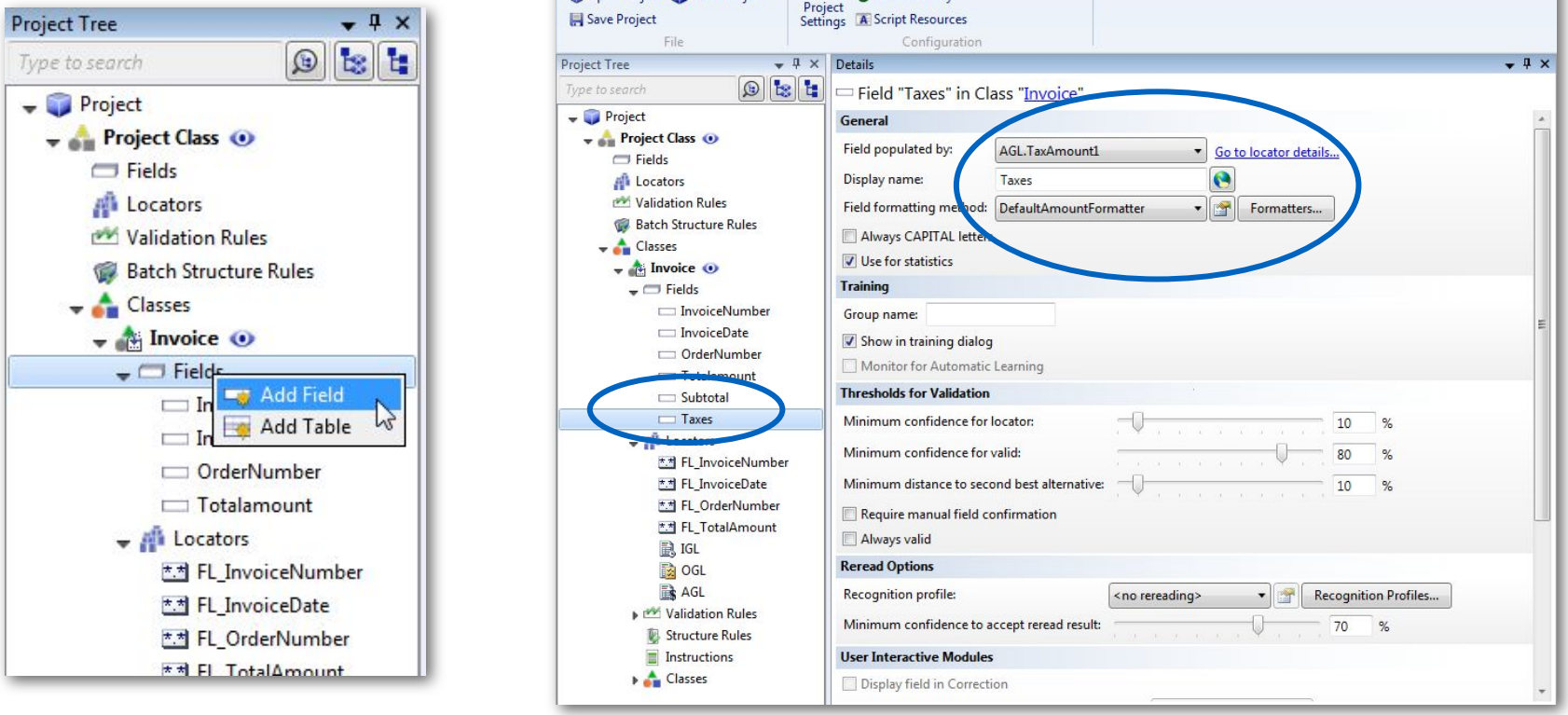
Затем давайте попробуем изменить вход локатора для четырех созданных нами полей. IGL вернет номер счета и дату, OGL вернет номер заказа. И AGL вернет общую сумму.

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



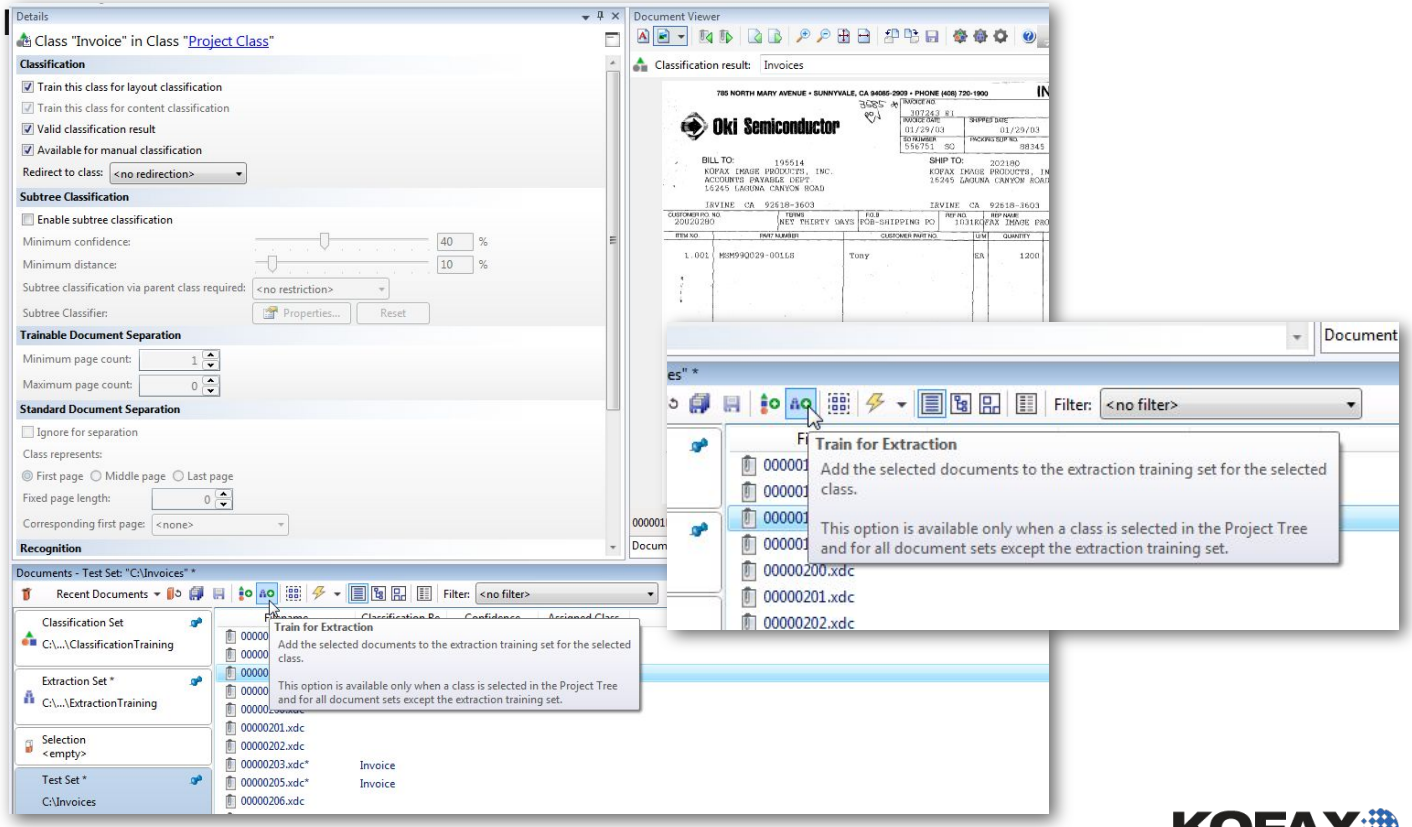
Новые поля

- 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 форм

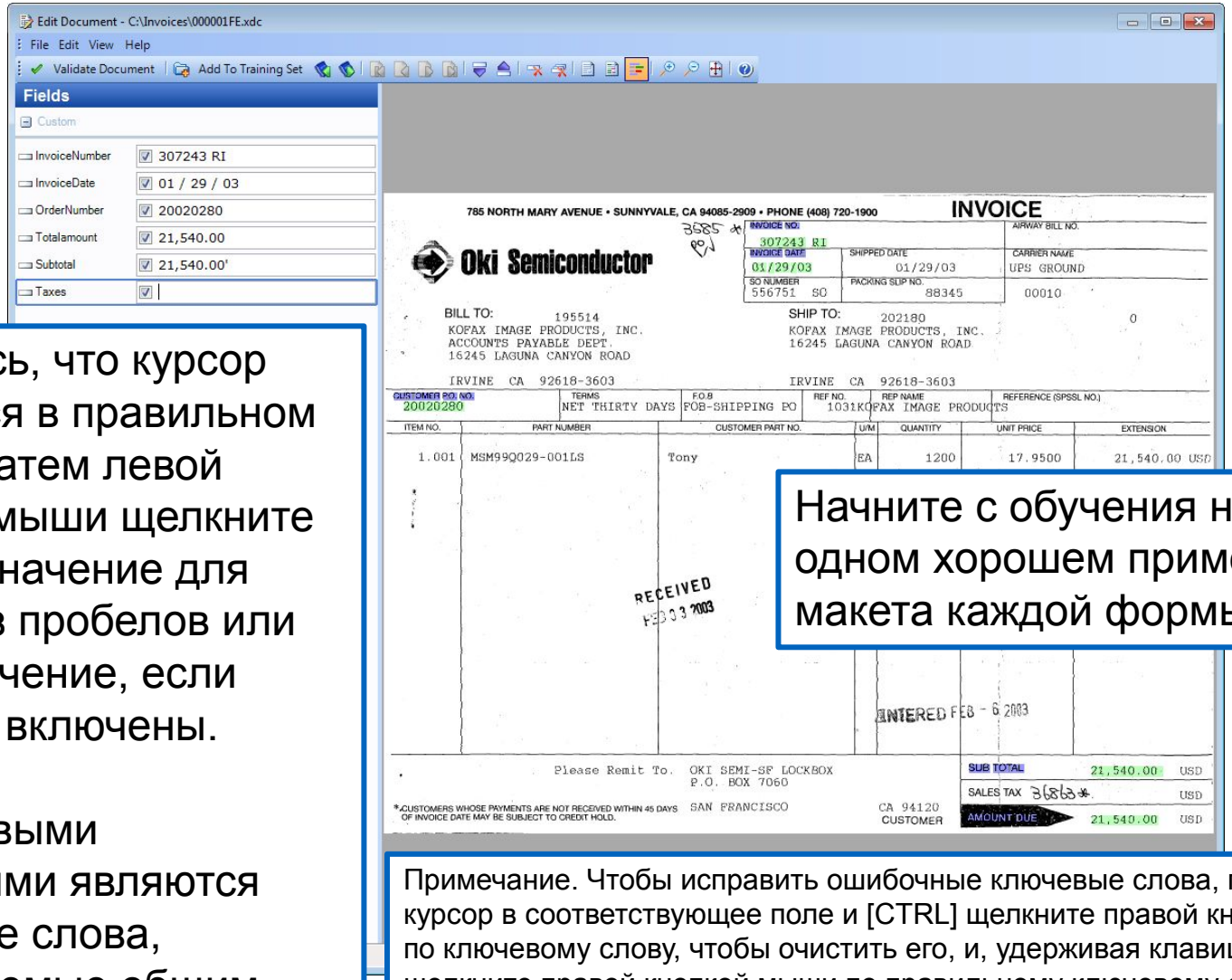


Теперь нам нужно обучаться извлечению

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 согласованны



Пробуем в каждом поле левой кнопкой мыши или рисованием



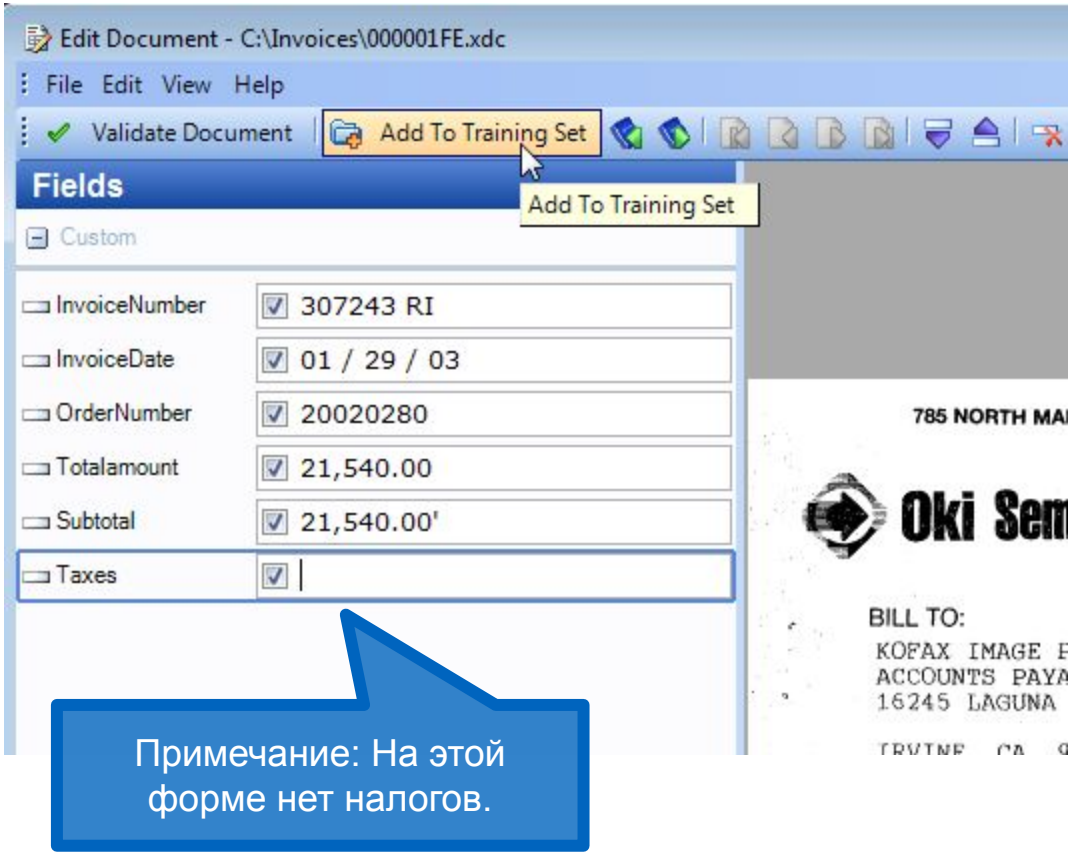
Убедитесь, что курсор находится в правильном поле, а затем левой кнопкой мыши щелкните нужное значение для строк без пробелов или lasso значение, если пробелы включены.

Фиолетовыми значениями являются ключевые слова, используемые общим алгоритмом для

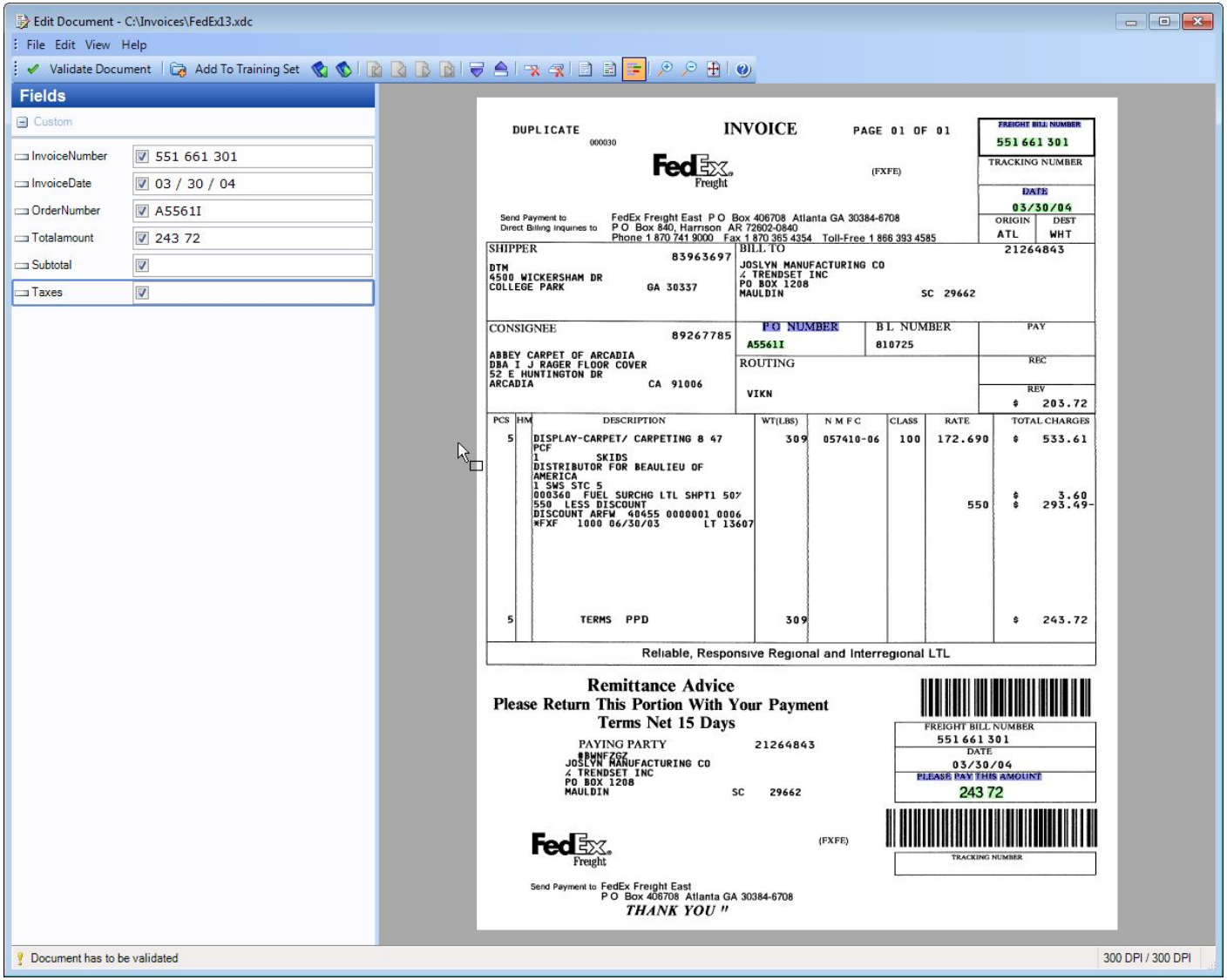
Начните с обучения на одном хорошем примере макета каждой формы.

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.

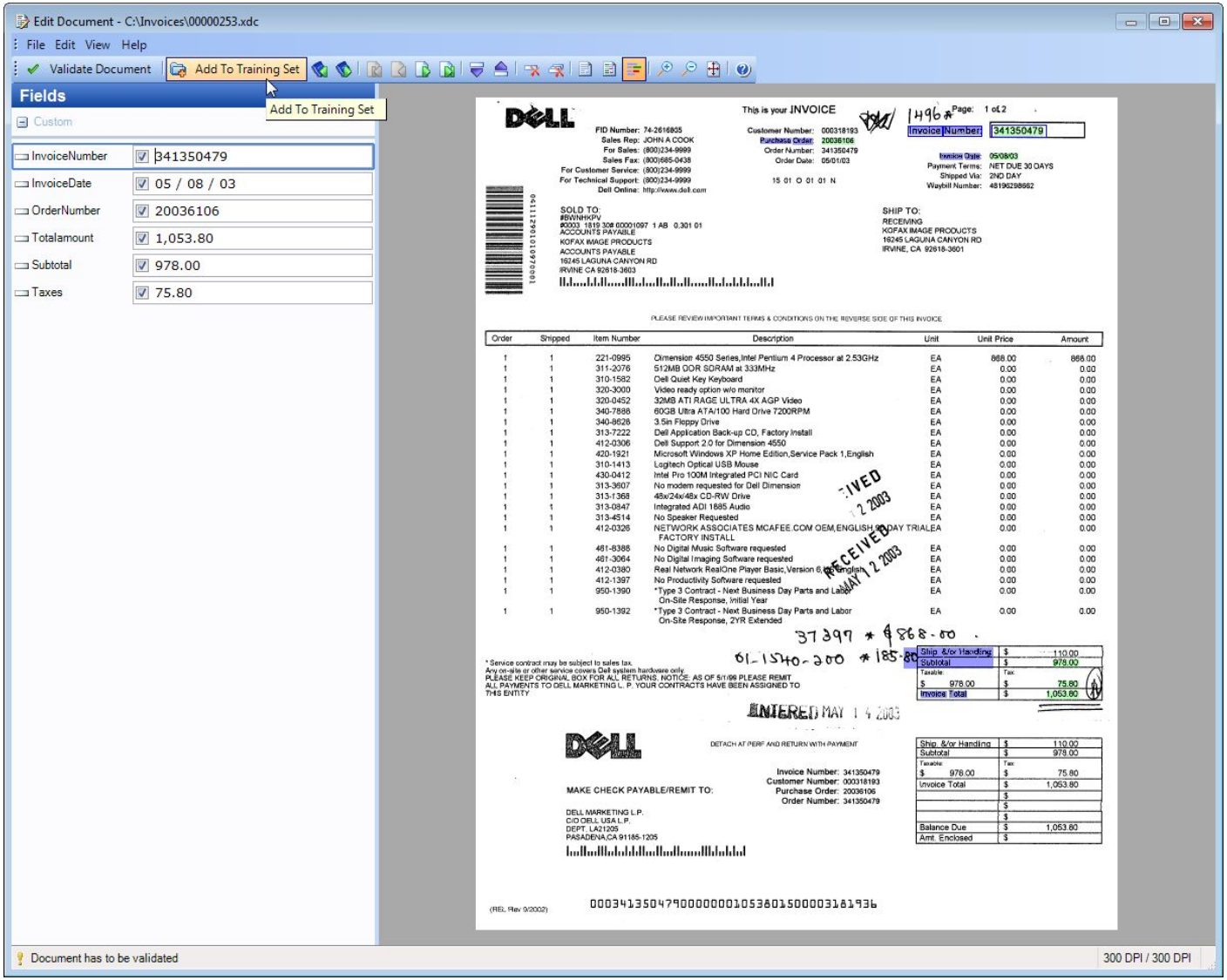
Добавить тренировочный комплект



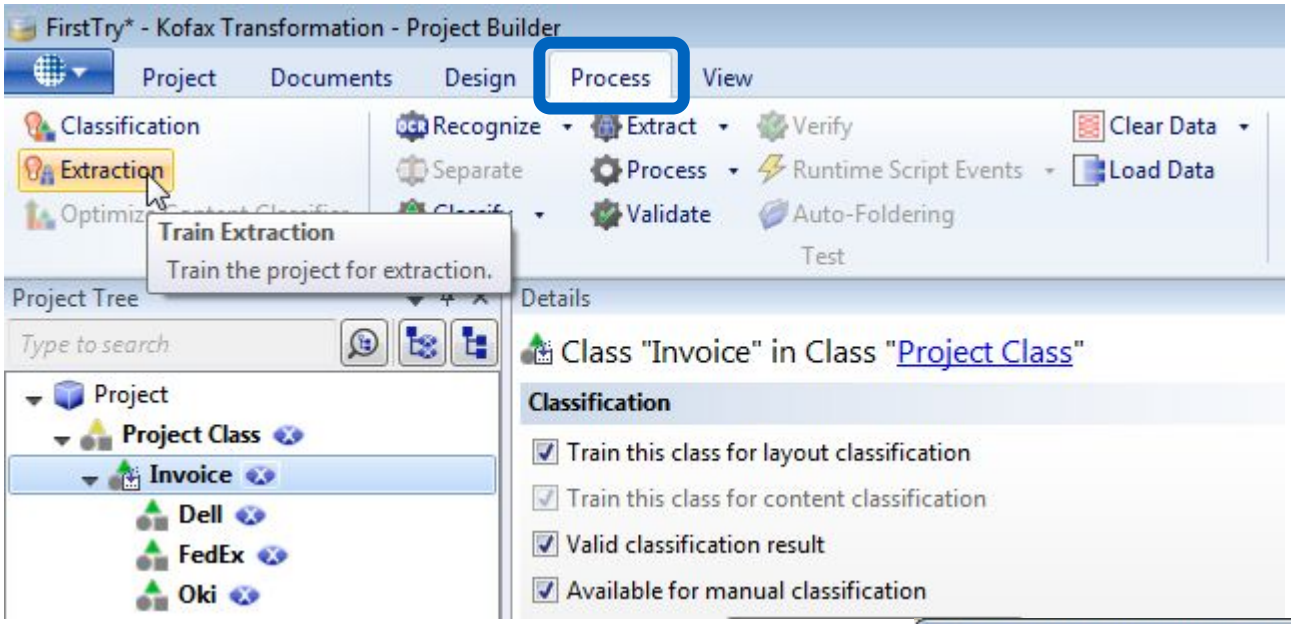
Обучаемся на других документах



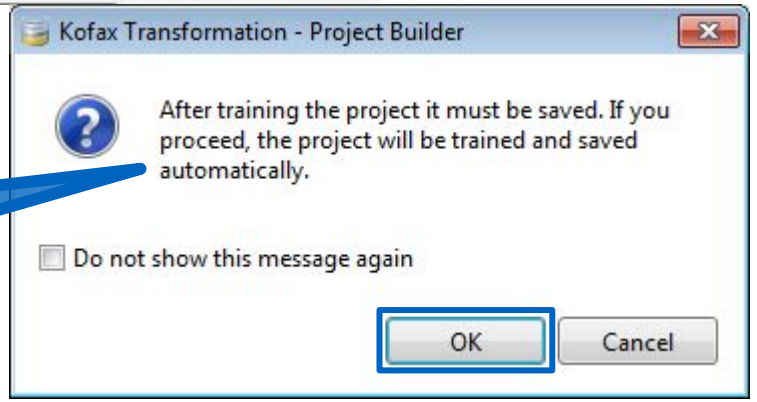
Добавляем другой



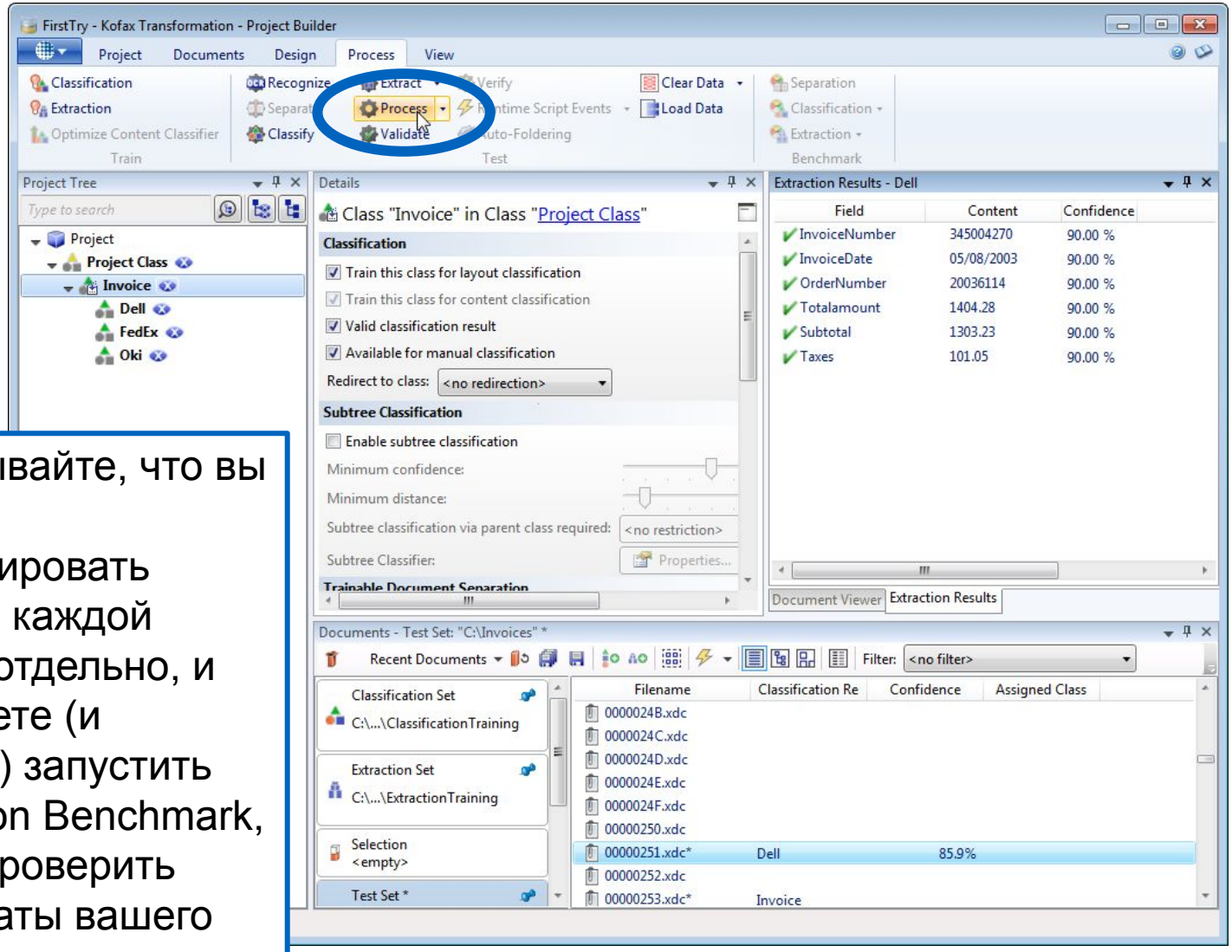
Обучаемся на проекте



Автосохранение. Спасибо, Project Builder!



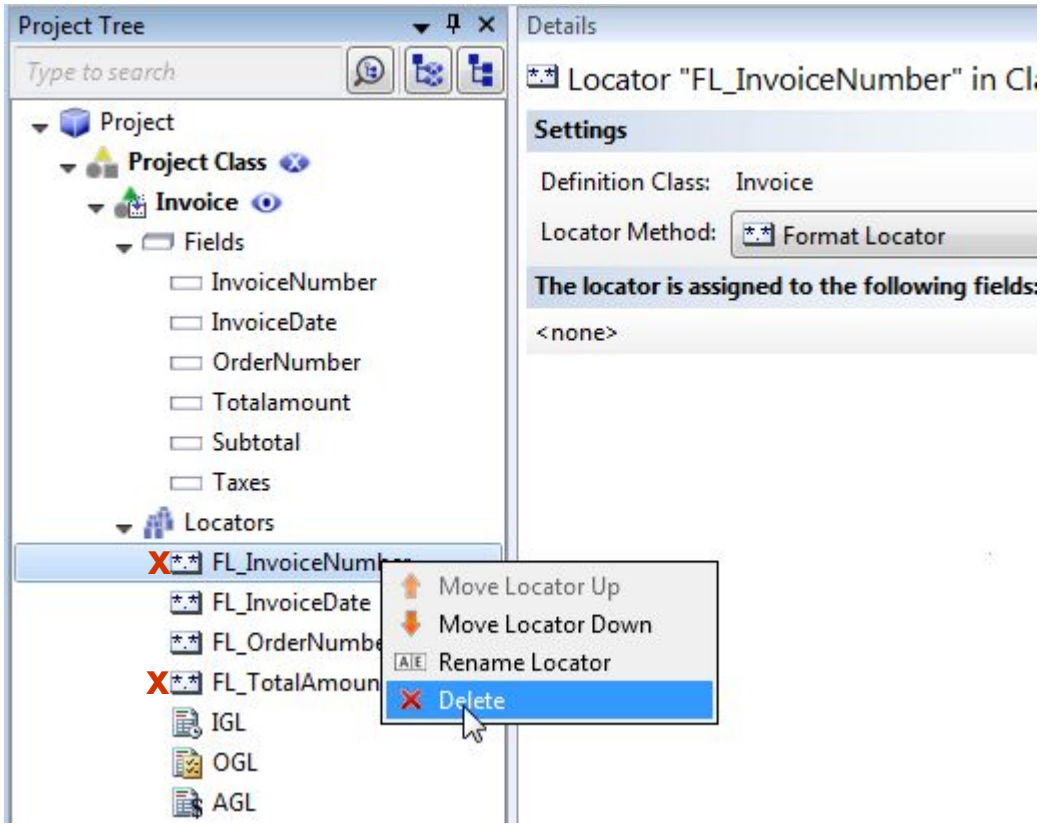
И тестируем



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

Удаляем неиспользованные 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...

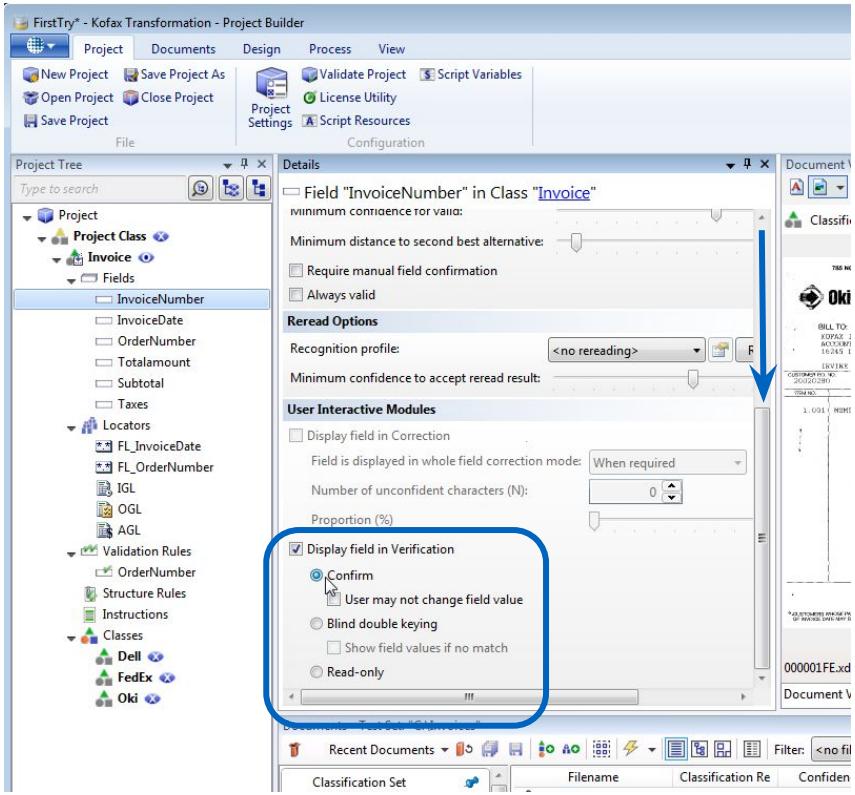


Добавляем KTM Верификацию

- 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.

- 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.

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



Демонстрация и задание