Kofax Transformation Modules Generic Versus Specific Online Learning

Application Note

Date	June 27, 2011
Applies To	Kofax Transformation Modules 3.5, 4.0, 4.5, 5.0
Summary	This application note provides information about Generic Versus Specific Online Learning
Revision	1.0

Kofax Transformation Modules Generic Versus Specific Online Learning Overview

This document describes the differences between Generic and Specific Learning with Kofax Transformation Modules, when each should be used, the effects on performance, and some common issues along with ways to resolve them.

What is Specific and Generic Learning?

Generic and Specific Learning are features of the Invoice Group Locators and the Trainable Group Locator. A Group Locator can be configured to use either Generic or Specific Learning or both. For Generic Learning, the learnt information is held in a Generic Knowledge Base. For Specific Learning, the learnt information is held in a Specific Knowledge Base. Generic Learning can be applied to single data fields, whereas Specific Learning can be applied to single data fields or tables. Generic and Specific learning can be applied to semi-structured documents such as invoices, sales orders, bills of lading etc. In general, a semi-structured document is one where:

1. There are multiple but repeated layouts
2. There are common data fields
3. The format of the document is outside the receiver's control

For a given semi-structured document type, Generic Learning is designed to achieve a good level of extraction performance (typically around 70%) over a wide range of layouts after training on 50-100 samples of varying layout. Specific Learning is designed to achieve a high level of extraction performance (typically > 90%) for given layouts after training on 1-3 samples.

> **NOTE:** Generic Learning never achieves higher that 70-75% recognition.

Specific Learning is able to give a high level of extraction performance, typically around 90-95% for each vendor, after training on 1-3 good samples. This substantially reduces the cost of configuring a solution for high extraction performance on the most common layouts. Generic Learning can still be used, but is usually most appropriate for the less common layouts.

Differences between Generic and Specific Learning:

The following table summarizes the differences:

	Generic Learning	Specific Learning
Usage	Used for documents of unknown layout.	Used for documents of a known layout.
Layout dependency	Based on a single model for the syntax of all documents and learns the differences between different layouts.	Layout specific – Based on one model for each different layout.
Training requirements- effort to deploy	Trained on many samples – typically requires 50-100 samples for good recognition on unseen layouts.	Trained on few samples – typically requires only 3 samples for good recognition from a particular layout.
Typical recognition rate	Recognition rate – Typically 70% over all layouts after sufficient training.	Recognition rate – Typically 90-95% for each layout after sufficient training.
Level of difficulty- skills required	Moderate difficulty – IT skills needed.	Simple – end-users can do it.

Specific Online Learning:

Specific Learning should be used to get high extraction performance for the most common layouts, i.e., the subset of all layouts that make up 50% of the total volume. The learning should be performed before the system is put into production. While running in production, Specific Learning should be used to train on the next most common layouts. Specific Learning remembers the layout of trained documents. During extraction, layout classification is performed. If the layout is known, then the document is extracted using information derived from the trained document of the same layout. If the layout is not known, then nothing is extracted. So, for Specific Learning to work on documents of a particular layout, at least one document of that layout must have been trained.

A Group Locator can be configured to use both Specific and Generic Learning from within Project Builder. In this case, if extraction using Specific Learning fails because the document layout is not known, then the Locator will then use Generic Learning to extract the data.

Specific Online Learning Requirements, Setup, and Notes:

1. Specific Learning needs to be trained on 3 documents of each layout to give good extraction results for that particular layout. Typically, no additional benefit is seen if more than 3 are trained. Specific Learning models each layout separately, whereas Generic Learning adds all layouts to a single model.

2. In order to build a Specific Knowledge Base, please use Project Builder. In Project Builder, select documents stored on disk or sent back from the Validation module. Train the documents and then

optionally create a Knowledge Base. Add this to the project, resynchronize and then republish the batch class.

3. Online Learning works differently with Specific Knowledge Bases than Generic Knowledge Bases. When a document is reviewed in Validation, a button can be pressed to mark it for Specific Online Learning. At the end of the batch, all marked documents are sent to the Knowledge Base Learning Server – this creates a new Dynamic Specific Knowledge Base that is available for use by the Extraction Server for the next batch (assuming Online Learning is enabled for the project). The documents can also be loaded into Project Builder where a new Published Specific Knowledge Base can be created off-line and then deployed into the live system. If the ‘Enable Online Learning’ option is enabled for the project, then the benefits of Specific Learning are seen in the next batch. All documents that were marked for Specific Learning by the user in KTM Validation are processed by the Knowledge Base Learning Server, which creates a new Knowledge Base in a few seconds. This Knowledge Base is then picked up by the KTM Server module at the start of the next batch. If the ‘Enable Online Learning’ option is not enabled for the project, then the benefit of Specific Learning cannot seen until the training documents sent back from Validation have been reviewed in Project Builder or Knowledge Base Administration, and the project has been retrained and resynchronized with the KC batch class and the batch class republished. The updated project and Knowledge Base is then picked up by the KTM Server module at the start of the next batch.

NOTE: Once the documents have been imported into Project Builder, the corresponding Dynamic Specific Knowledge Base is no longer available in production. The project must be retrained (or a new Knowledge Base created) in Project Builder and the project resynchronized and republished. Online Learning will work with KCNS (formerly ACIS) only if the Knowledge Base Learning Server is run on the same LAN as the extraction server (because the two modules share files). Documents marked for Online Learning in KTM Validation at the Remote Site are sent back to the Central Site as part of the batch.

4. Online Learning can be disabled at the project level. If Online Learning is enabled, then documents marked for Specific Online Learning in KTM Validation are used by the Knowledge Base Learning Server to create a new Dynamic Specific Knowledge Base at the end of the batch. If Online Learning is disabled, then documents marked for Specific Online Learning in KTM Validation can be loaded into Project Builder, from where a new Published Specific Knowledge Base can be created off-line and deployed into the live system. The capability to disable Online Learning is provided to cater for end-users who do not want their system to learn in production.

5. A document can be marked for both Specific and Generic Online Learning in the KTM Validation module. Documents marked for Specific Learning are used by the Knowledge Base Learning Server to build a Specific Knowledge Base on-line and can also be used in Project Builder to build a Specific Knowledge Base off-line.

6. It is not recommended to use Specific Knowledge Bases that were developed for one environment with a different environment. In reality, each customer has a different set of common layouts so one customer’s Specific Knowledge Base will be of little use for another. However, remember that Specific Knowledge Bases are quick to train, so developing one for a customer as part of a Proof of Concept (POC) or during deployment will be relatively quick and cheap.

7. When a user marks a document for Specific Learning in the KTM Validation Module, a document of the same layout in a subsequent batch will always be seen in KTM Validation regardless of whether the data is valid or not. If the user confirms that the data is extracted correctly for this layout, and then the next time a document of the same layout is encountered, assuming all data is valid, the user will not see it in KTM Validation again. If the user has to amend the data in

some way, e.g., because the training first time around was incorrect, then they should mark the document for Specific Learning again. On the next pass, if the user confirms that the data is extracted correctly for this layout, then the next time a document of that layout is encountered, assuming all data is valid, the user will not see it in KTM Validation again. Essentially, the Specific Learning is looking for consistency in the training – the more consistency it sees the more confident the recognition becomes. This procedure is essentially a safety check that prevents badly trained documents from reducing data extraction accuracy.

Reasons to Create a SKB (Specific Knowledge Base):

Some reasons to create a new, published Specific Knowledge Base off-line in Project Builder are the following:

1. It consolidates all the “knowledge” in the project into one place, thus making it easier to take a system snapshot that can act as a roll-back point.
2. There are some slight benefits in processing speed to be gained.

A limit to the number of documents that can be added to a dynamic Specific Knowledge Base can be set for the project; once this limit is reached, no further learning can be performed on-line. At this point the documents can be loaded into Project Builder from where a published Specific Knowledge Base can be created using all documents or just a subset and redeployed into the production system. Online Learning can then be performed until the limit is reached again.

Generic Online Learning:

Generic Learning should be used to get good extraction results from the remaining less common layouts.

For a given semi-structured document type, Generic Learning is designed to achieve a good level of extraction performance (typically around 70%) over a wide range of layouts after training on 50-100 samples of varying layout.

In contrast to Specific Learning, Generic Learning works irrespective of layout, so in order for Generic Learning to work on a document of a particular layout, it does not necessarily have to have been trained on that layout.

A Group Locator can be configured to use both Specific and Generic Learning from within Project Builder. In this case, if extraction using Specific Learning fails because the document layout is not known, then the Locator will then use Generic Learning to extract the data.

Generic Online Learning Requirements, Setup, and Notes:

1. 50-100 samples of varying layouts per class.
2. In general, a semi-structured document is one where:
   a. There are multiple but repeated layouts.
   b. There are common data fields.
   c. The format of the document is outside of the receiver’s control.
3. Generic Learning is mainly based on keywords and their location in relation to an alternative or field. As a result, this option requires good OCR quality in order for data to be extracted from documents. Generic training is also able to extract values from unknown documents that use similar keywords as some of the trained sample documents.

4. In contrast to Specific Learning, Generic Learning should only be trained on a document of each layout, but there is a lower probability that another document of that layout will be extracted confidently. Specific Learning models each layout separately, whereas Generic Learning adds all layouts to a single model.
5. Documents marked for Generic Learning can be used to build a Generic Knowledge Base off-line. The benefit of Generic Learning is not seen until the training documents sent back from Validation have been reviewed in Project Builder or Knowledge Base Administration, and the project has been retrained in Project Builder, resynchronized with the Kofax Capture batch class and that batch class has been republished. The updated project and Knowledge Base is then picked up by the KTM Server module at the start of the next batch.

Reasons to Create a GKB (Generic Knowledge Base):

1. Documents that were not properly extracted are used to improve the extraction results for your project through a GKB. This training is typically the responsibility of the system administrator who processes sample documents that have been placed in a training set. The training session creates new extraction patterns that are stored with the project.

2. Portable binary representation of trained data.

Common Problems/Issues:

1. **Knowledge Base Learning Server Fails for Duplicate Image File Names. An error message, “1_x.tif already exists,” is displayed if the Knowledge Base Learning Server attempts to save an image with a duplicate file name to the directory.**

   This problem has been resolved. All images are given unique names when processed by the Knowledge Base Learning Server, eliminating the possibility of duplicate file names (SPR00065586). This issue has been resolved in KTM 4.5 and 5.0. Please upgrade to the latest Service Pack and Fix Pack for KTM 4.5 or 5.0.

2. **KTM Server and Knowledge Base Learning Server Error, FOX library error message, “Could not Create Lock File,” may be displayed when all extraction services try to use the learning database at the same time.**

   This can happen if Server processes large batches and many documents for Online Learning.

   This problem has been resolved (SPR00051589). Please upgrade to the latest Service Pack/Fix Pack/ Patch for KTM 4.0 or 4.5.

3. **I am using a Trainable Group Locator (TGL) in Kofax Transformation Modules (KTM). It has been extracting fine, with Specific Online Learning in use. I added a new field to this locator, and the Specific Online learning for the locator does not work and extraction is not occurring as desired. How can I resolve this issue?**

   The confidence threshold for the locator was set too high. Try to reduce it to a lower threshold, such as 30 or 40, and test. Once extraction and Specific Online Learning are functioning again, slowly increase the confidence threshold to see what threshold is appropriate for this particular project.

4. **How can I resolve the error in KTM Server "BeforeExtraction: BeforeExtract: the execution of a locator method failed.Class = 'Invoice' , Locator = "LOCPO", Original error message: Unknown error in FoX library occured. This file was saved by a newer version of the FoX image classifier"**

   This problem may be due to a corrupt Dynamic Knowledge Base. To resolve, stop all KTM services, set all Online Learning to a static Knowledge Base and then disable 'Online Learning'. Once disabled and then re-enabled, it recreates the online learning files. When extraction is restarted, the issue should be resolved.

5. **What causes the below error when processing batches through KTM Server, using a KTM project that has specific online learning enabled?**
   "BeforeExtraction: BeforeExtract: The server threw an exception. (Exception from HRESULT: 0x80010105 (RPC_E_SERVERFAULT))"

   If you have many specific sample documents, it is possible that those contain some corrupted documents. Usually, this can be solved by creating a Specific Knowledge Base from the specific samples. Perform the following steps:

   1. Open the KTM project in Project Builder.
   2. Open the specific samples, and select the "Import documents from Specific Online Learning" option. This will create a new training folder.
   3. Open Project Builder's Project Settings, and on the Knowledge Base tab you can create a Specific Knowledge Base from this new training folder.

6. **I am running the Knowledge Base Learning Server (KBLS), and receive the following error message: "Failed to load project, because: This file was saved by a newer version of the "Specific Knowledge Base Algorithm" extraction method". How can I resolve this issue?**

   Two possible causes are:

   - Corrupt KTM Project. This may have been caused by copying project files/folders.

     A recommended step for a similar or exact iteration of a project is to perform a *Save project as,* with the copy of the current project being named differently. It may also be recommended to have a separate project folder as well.

   - Corrupt Online Learning Directory.

     This can be resolved by creating a new Online Learning Directory (QAID 13215).

   It is recommended at this point to revert to a previous backup of a project, or to another working iteration of the copied project.

7. **Are Dynamic Specific Knowledge Bases portable?**

   No, Dynamic Specific Knowledge Bases are not meant to be portable. To move the training included in Dynamic Specific Knowledge Bases, they should be merged with the project.

   In Project Builder, select *New Samples* in the lower pane, click *Display Specific New Samples,* and then click *Import Documents from Specific Online Learning.* This merges the current generation of Dynamic Specific Knowledge Bases with the project.

Once this has been merged with the project, a Knowledge Base can be created from the *Knowledge Bases* tab in the locator's properties (KTM 4.x) or the project settings (prior to KTM 4.0) or the Knowledge Base Administrator.

The created Knowledge Base is a portable file which can be moved to another system.

For more details, please see the Configuration Guide for the appropriate version of KTM.

8. How can I change the Online Learning directory?

KTM 3.5

1. Open KTM Project in Project Builder.
2. Open *Project Settings* and go to the *Knowledge Base* tab.
3. Check (set/enable) the *Enable Online Learning* option, if not already enabled.
4. Change the Online Learning directory by editing the path to a location that has never before contained Online Learning files.

KTM 4.0, 4.5, 5.0, 5.5

1. Open KTM Project in Project Builder.
2. Open *Project Settings* and go to the *General* tab.
3. Check (set/enable) the *Enable Online Learning* option, if not already enabled.
4. Change the Online Learning directory by editing the path to a location that has never before contained Online Learning files.

As a suggestion, naming the new Online Learning directory to the project name, and possibly even the date changed, may assist later in determining the Online Learning directories for particular projects and when they were changed.

NOTE: Changing the Online Learning directory recreates all Online Learning files.

9. What is the maximum number of documents that can be used for the Online Learning feature?

- The maximum number of documents that can be stored for input when using the Online Learning feature is 20,000.
- The default value is set at 2,000.
- The minimum number of documents can be set at 100.

10. I am trying to publish a Batch Class that uses KTM, and receive an error message: "Configured online learning path not available". How can I resolve this issue?

- Check if the Online Learning path is actually correct, and that the currently logged in user has the correct permissions to access this path.

  If the above is correct, then the path to the specified project is incorrect. It is very easy to specify incorrect paths, especially if there are multiple local copies of a project on a machine.

- Double-check the path reference in Project Builder and ensure that you are using the same path in the KTM Synchronization Tool.

11. **After a Batch is processed by the Knowledge Base Learning Server, why are there no samples located in the path specified for Online Learning?**

    This can occur if the Knowledge Base Learning Server is used in a workflow with a project that ***does not use any Group Locators***.

    The KB Learning Server is intended to help improve the accuracy of Group Locators (*Amount, Order,* or *Invoice* Group Locators).

12. **How can I set up Extraction or the Knowledge Base Learning Server as a service?**

    The Kofax Transformation Modules - Server and Knowledge Base Learning Server can either run as an application, where it has a graphical user interface, or it can run in the background as a Windows service. In either case, document processing such as classification, extraction, and folder processing for the Server or Online Learning for the Knowledge Base Learning Server is directed to a Windows service.

    Kofax Transformation - Service Configuration

    The service configuration allows changing initial settings made when Kofax Transformation Modules was installed. The following settings can be modified:

    - Log on account for the services
    - Service startup type
    - User Profiles for Kofax Capture Administration

    To start the Service Configuration

    1. Select Kofax Transformation ¦ Service Configuration from the Windows Start button. For example, for a Windows XP system: Start ¦ All Programs ¦ Kofax Transformation ¦ Service Configuration. The Kofax Transformation - Service Configuration dialog box opens.

       NOTE: This sequence may differ according to the version of the operating system.

    2. Change the settings. For further details, see Kofax Transformation Modules - Server and Services online help topics.