Maintaining Extraction Online Learning

Application Note

Date	October 4, 2019
Applies To	Kofax Transformation Modules
Summary	This application note goes into detail about best practices for maintaining Online Learning in a KTM Project.
Revision	2.0

Online Learning Overview

Online Learning is a method of using unsuccessful classification or extraction results to improve documents processed in the future. If a document is not correctly classified or extracted during production, once the Validation User corrects the unsuccessful classification or extraction results the Online Learning server can collect the document and apply it to the project document training sets for improved classification and extraction results for future extraction. These documents can be marked for Online Learning by Validation Users, flagged automatically by KTM, or through script. This guide will specifically go over how to maintain Extraction Online Learning.

Documents collected for Extraction Online Learning are used to improve extraction for Trainable Locators. These locators use the geometric information of documents with the same layout to extract data. Without field position and coordinates, the specific training algorithm cannot train the documents correctly. The following locators fall under this category:

- Trainable Group Locator (TGL)
- Amount Group Locator (AGL)
- Invoice Group Locator (IGL)
- Order Group Locator (OGL)
- Table Locator (If training is enabled for the locator)
- Line Item Matching Locator (LIMLoc)
- Text Content Locator
- Trainable Evaluator

These documents are collected and stored into the Online Learning directory. To apply this learning, the documents need to be imported from the Online Learning directory into the Extraction Set of the KTM Project. It is recommended to import these documents into the Extraction Set when the Online Learning directory is close to reaching the default limit of 2,000 documents. Depending on how quickly the Online Learning directory is collecting documents an import may be needed as often as every two weeks or as little as every two months. This guide will demonstrate how to properly import Online Learning Samples into the KTM Project.

Ideally there should be a point where all types of documents have been learned on using Online Learning and the feature can be turned off. However there are some KTM projects that will continue to get different type of forms layout and Online Learning is never turned off. With projects such as these it is vital that online learning is properly maintained to prevent the project from growing to unwanted sizes.

Once an Extraction Set is properly trained, the documents in the set can be compressed into knowledge bases. A knowledge base in KTM is a binary file used to store extraction patterns. Knowledge bases are relatively compact. The TGL, AGL, IGL, OGL, and Table Locator can all utilize knowledge bases for extraction. There are three more locators that use online learning: Line Item Matching Locator (LIMLoc),

Text Content Locator, and the Trainable Evaluator. These three locators are not covered in this guide because they use their own online learning models and do not use the Extraction Set or knowledge bases.

Knowledge bases are the key to proper Online Learning maintenance. It is better to have a knowledgebase with a learning model of 5,000 invoices rather than keeping 5,000 invoices because it helps maintain the size of the project and increase efficiency of the project.

There are two practices for creating knowledge bases:

- Incremental: A knowledge base is created every time Online Learning documents are imported. The newly created knowledge base only consist of the recently imported documents. For example after one year of Online Learning maintenance, each locator could have 4 knowledge bases assigned to it (one created for every 3 month import of documents).
- Cumulative: A knowledge base is created every time Online Learning documents are imported. The newly created knowledge base uses extraction data from all documents in the Extraction Set. Every document used for training is saved in the Extraction Set since they will be needed for every new knowledge base created.

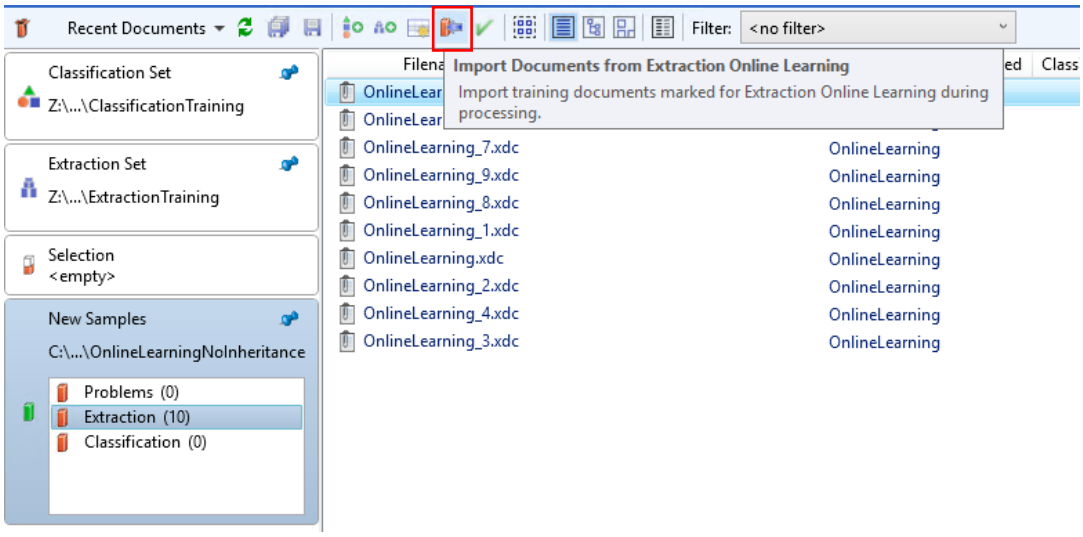
Note that there is no specific setting that automatically creates an incremental or cumulative knowledge base. A knowledge base is considered incremental or cumulative based on the documents used to create the knowledge base. There is no built-in way to tell whether a knowledge base is incremental or cumulative unless specified in the knowledge base name. Ensure all users who are creating knowledge bases for a specific project are using the same practices.

Incremental knowledge bases take less time to maintain and keeps the KTM Project/Extraction Set to a reasonable size, but can provide less accurate extraction results. This is because trainable locators only pull the results from one knowledge base for each document extracted. Cumulative knowledge bases result in better extraction, but can be difficult to maintain if there are too many documents collected in shorts amount of time.

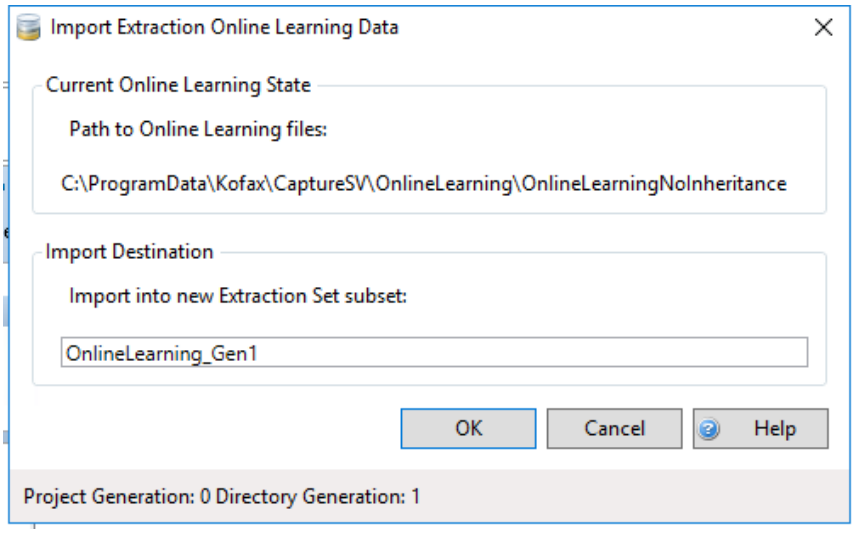
With incremental learning all documents in the Extraction Set are deleted after maintenance is done, but with cumulative learning most of the documents are not deleted. If there is an issue with a previously created incremental KB, there is no way to edit and recreate the KB since the original documents are deleted. Depending on how quickly documents are collected, cumulative learning can lead to a large and difficult to maintain Extraction Set. The larger the extraction set grows, the longer it takes to train documents and resolve conflicts. If there are thousands of documents being collected in a short amount of time and Online Learning is never turned off, cumulative learning is not recommended.

Online Learning Maintenance Steps

1. Open the KTM Project in Project Builder, navigate to the Document Viewer and click on the “New Samples” Document Set. After selecting the Document Set click on the button “Import Documents from Extraction Online Learning” to start the import process.

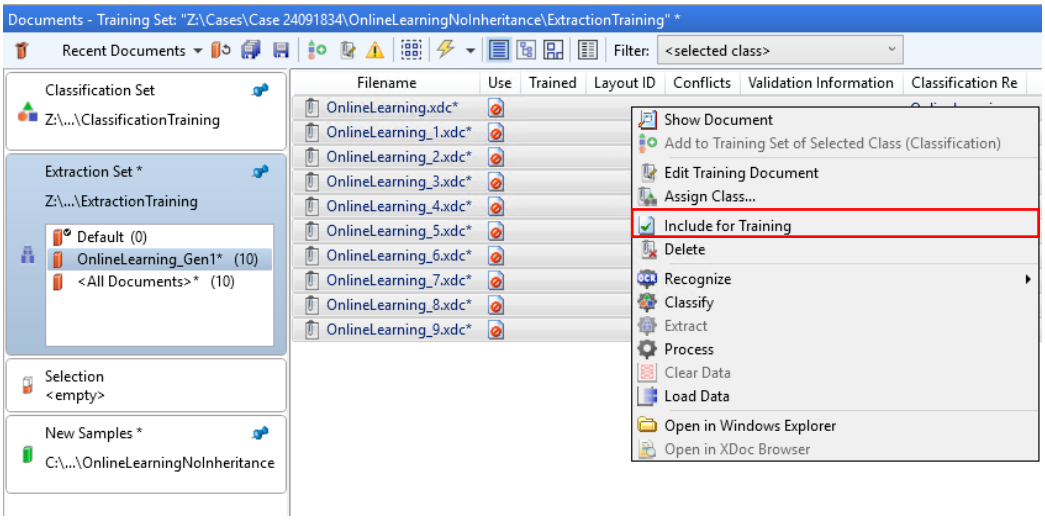


2. The Import Extraction Online Learning Data Window is used to import online learning documents in the New Samples Set into the Extraction Set. The New Samples will be imported as a new Subset into the Extraction Set. The name of the subset can be customized, but by default it is “OnlineLearning_Gen<number>” where number is the latest Online Learning generation number (as demonstrated below). Press OK to start the import.

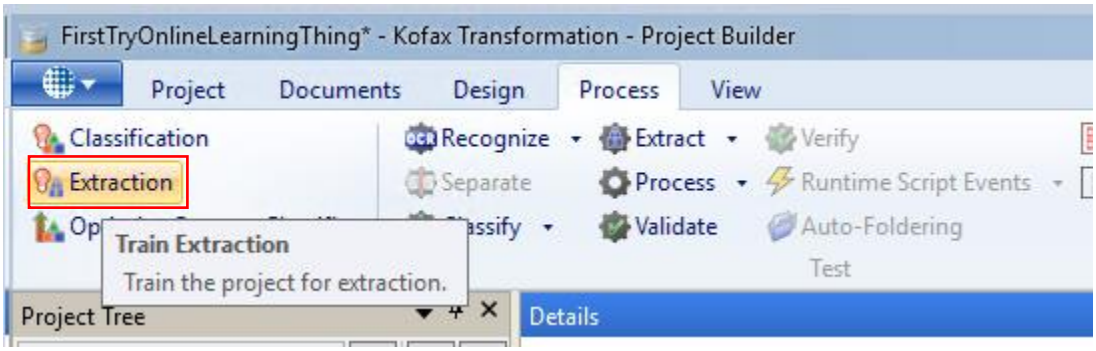


3. Check the Extraction Set to make sure the Subset was created successfully.
    - If practicing incremental knowledge base creation, select the set <All Documents>, highlight all documents, right click and select “Include for Training”.
    - If practicing cumulative knowledge base creation, select the set <All Documents>, highlight all documents, right click and select “Exclude from Training”. Then go to the newly created subset, highlight all documents, right click and select “Include for Training”.

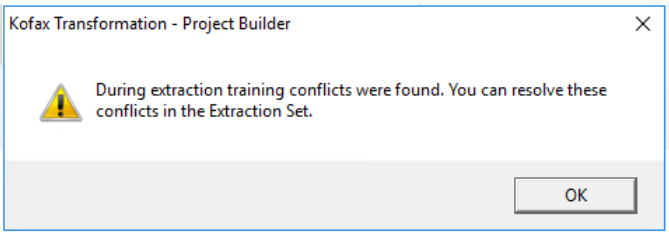
    When this is done the Use column should go from all red to all green (Use column is red/excluded in the screenshot below)



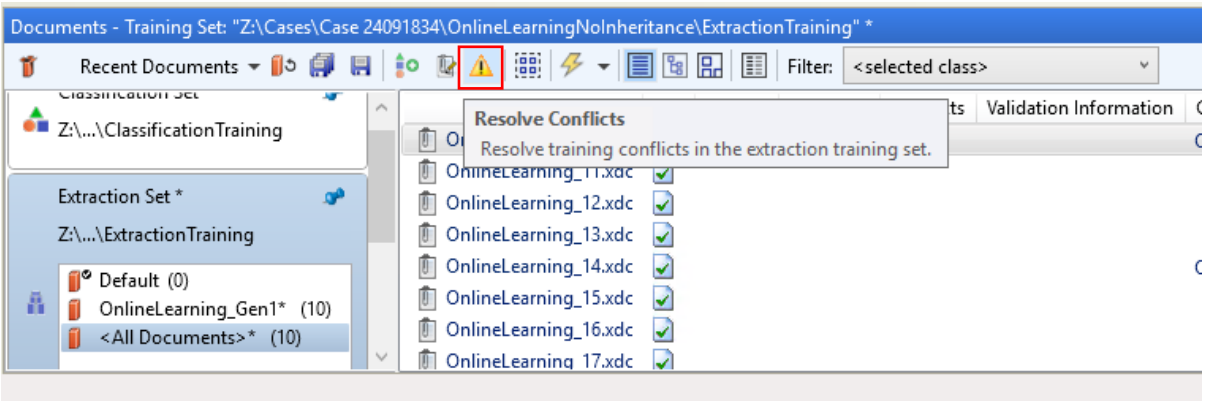
4. In the Project Builder Ribbon Bar click the tab “Process” and then click “Extraction” to train the project for Extraction using the newly imported documents. There will be a prompt about the Extraction Set changes being saved. Click OK on this prompt.



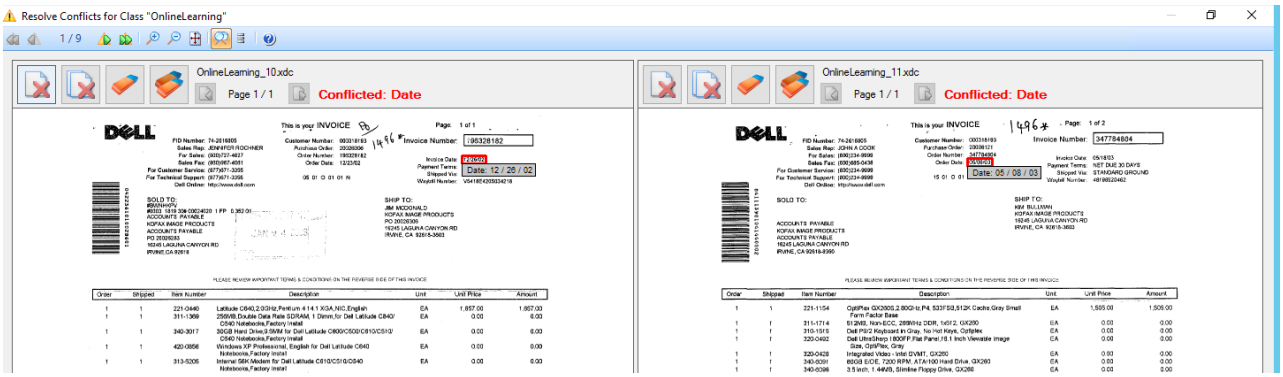
5. After training the following prompt may appear. This indicates that while training the project for extraction there was conflict data that needs to be resolved.



6. In the Document Viewer go back to the Extraction Set and click “Resolve Conflicts”. A prompt will show up about the document set being saved automatically. Click OK to this prompt.



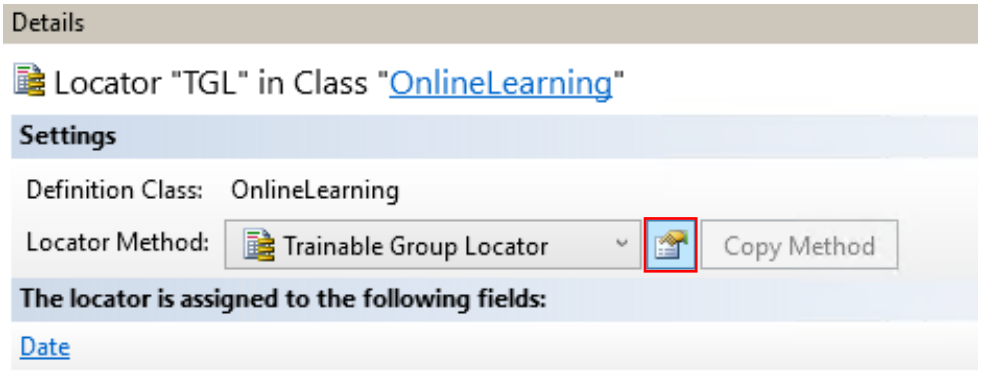
7. The Resolve Conflicts window will appear and show the user all existing conflicts in the Extraction Set. In the example below, the conflict is that the field “Date” was trained in two different locations. The conflict can be resolved by doing the following:
   - Delete the field data in one of the documents
   - Delete the field data in both of the documents
   - Delete one of the documents
   - Delete both documents



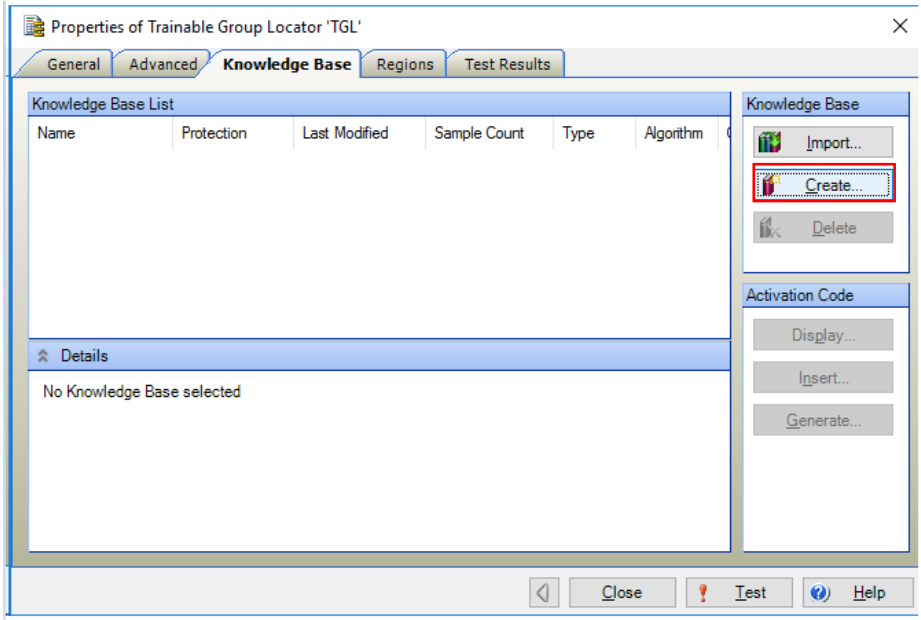
Conflicts arise when Validation Users validate the same field in different areas. This causes a discrepancy between the documents. Ideally Validation Users should try to validate data in the exact same location to prevent conflicts.

Note that any bad samples such as poorly scanned images should be deleted from the Extraction Set. Training sets should contain clean images with correctly extracted data. Poorly scanned images with invalid extraction data will result in less accurate extraction at run-time.

8. Once there are no more conflicts in the Extraction Set go to the properties of your Trainable Group Locator by clicking the properties button next to the locator method.

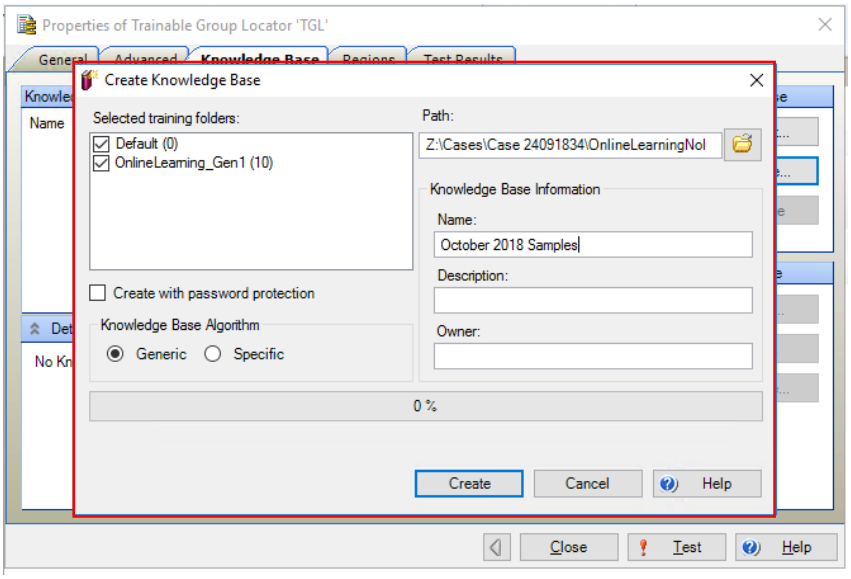


9. Navigate to the Knowledge Base Tab and then click Create…

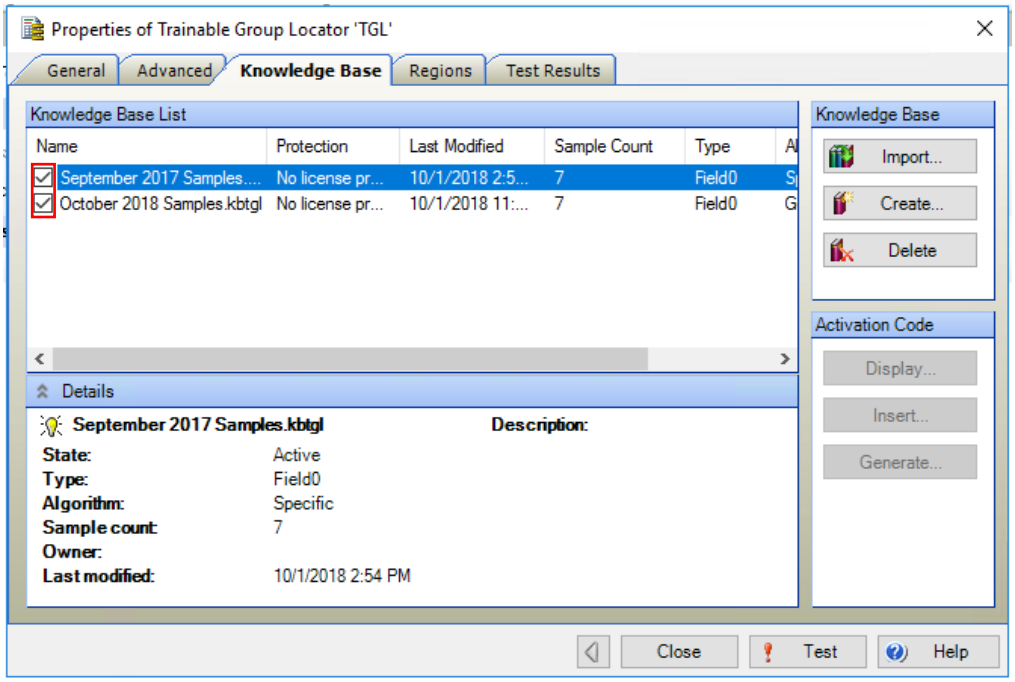


10. The Create Knowledge Base window is now displayed. Select the Knowledge Base Algorithm based on whether the Trainable Locators are configured for Generic or Specific learning. Generic learning is mainly based on keywords and keyword locations in a relation to an alternative or field. Specific learning will learn different layouts of documents and apply the extraction learning based on each document layout. Specific learning is the more popular approach, but it only works for trained layouts.
    - If practicing incremental knowledge base creation, make sure only the newly created subset is checked under “Selected training folders”. Provide a custom name to the knowledge base such as “October 2018 Samples” to label the samples.
    - If practicing cumulative knowledge base creation, make sure all folders are checked before creating the knowledge base.

    Once the configuration is complete click “Create” to create the new Knowledge Base.



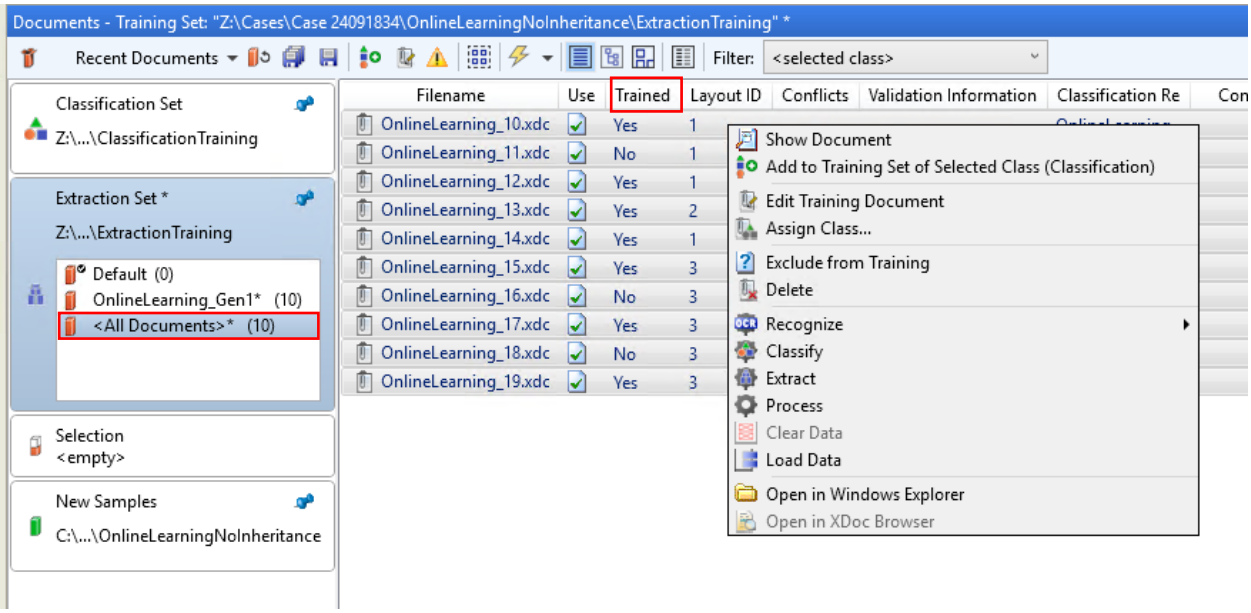
11. The new knowledge base should show up in the Knowledge Base tab for the locator. Make sure to check this knowledge base or else the locator will not use it.
    - If practicing incremental knowledge base creation, make sure all previous knowledge bases are also checked to use previously trained data (Image below shows incremental KBs for September and October).
    - If practicing cumulative knowledge base creation, make sure to uncheck all other knowledge bases. Only the latest KB is needed since it contains the data from all training documents.



12. Repeat Steps 9 through 12 for all other trainable locators defined in the project. Here is a list of all Trainable Locators that allow for Knowledge Base Creation:
    - Trainable Group Locator
    - Amount Group Locator
    - Invoice Group Locator
    - Order Group Locator
    - Table Locator (If training is enabled for the locator)

13. After all knowledge bases are created the Extraction Sets:
    - If using incremental knowledge bases select the “<All Documents>” set in the Extraction Set, highlight all documents, right click and delete all documents by clicking “Delete”.
    - If using cumulative knowledge bases select the “<All Documents>” and sort by the “Trained” column. Delete all documents where the value of “Trained” is No. The KTM Extraction Algorithm has determined that these documents are not needed for Extraction Learning and can be safely deleted. All documents with a “Trained” value of Yes must NOT be deleted because they will be used again to create the next cumulative knowledge base.

      Note that the Extraction Set should be kept to a reasonable size or else the project may grow too large in size which can cause poor performance and will greatly increase the amount of time taken to perform online learning maintenance.


Documents - Training Set: "Z:\Cases\Case 24091834\OnlineLearningNoInheritance\ExtractionTraining" *
Recent Documents
Filter:
<selected class>
Classification Set
Z:\...\ClassificationTraining
Extraction Set *
Z:\...\ExtractionTraining
Default (0)
OnlineLearning_Gen1* (10)
<All Documents>* (10)
Selection
<empty>
New Samples
C:\...\OnlineLearningNoInheritance
Filename
Use
Trained
Layout ID
Conflicts
Validation Information
Classification Re
OnlineLearning_10.xdc Yes 1
OnlineLearning_11.xdc No 1
OnlineLearning_12.xdc Yes 1
OnlineLearning_13.xdc Yes 2
OnlineLearning_14.xdc Yes 1
OnlineLearning_15.xdc Yes 3
OnlineLearning_16.xdc No 3
OnlineLearning_17.xdc Yes 3
OnlineLearning_18.xdc No 3
OnlineLearning_19.xdc Yes 3
Show Document
Add to Training Set of Selected Class (Classification)
Edit Training Document
Assign Class...
Exclude from Training
Delete
Recognize
Classify
Extract
Process
Clear Data
Load Data
Open in Windows Explorer
Open in XDoc Browser

Online Learning Best Practices

1. Stick with a single Online Learning practice (incremental or cumulative): While practicing incremental or cumulative will depend on the number of documents processed, make sure that the knowledge bases created are consistent. If there are multiple users in charge of maintaining online learning make sure all users are creating the same type of knowledge bases.

2. Consistently maintaining online learning: The thirteen steps listed above should be routinely scheduled for the health of the project. Depending on how quickly documents are collected this could be as often as every two weeks or as little as every two months.

3. Keeping number of knowledge bases enabled per locator to a minimum: Incremental knowledge base practice make it easier to maintain and create knowledge bases. However while a locator can have multiple knowledge bases assigned to it, only one knowledge base will be used for extraction per document. Say a knowledge base from two years ago and the latest knowledge base have learning data on the same document layout. Because the locator finds a match with the older knowledge base it will only use that to extract data. If possible leave the number of knowledge bases used to a minimum by unchecking or deleting older knowledge bases.

Online Learning Bad Practices

While Online Learning can be a great feature that can help increase extraction results, if can cause issues with Extraction or even the entire project itself if not maintained properly. Here are examples of common bad practices:

1. Never cleaning the Extraction Set: This is the most common issue and can also cause the most problems. If Online Learning is going to be enabled permanently, then the Extraction Set must be cleaned out regularly. It is not recommended to keep more than a couple of thousands documents in any training set. The collection of thousands of documents will cause the KTM Project to grow to an exponential size which will make the project difficult to manage.

2. If the Extraction Set has too many samples (EX: 10,000+), it will take a great amount of time to train the project for Extraction and resolve conflicts. Having too many samples can cause out of memory errors when training the project.

3. Marking documents for Online Learning through script: In older versions of KTM built-in Online Learning monitoring did not exist. As a result many projects used scripting to mark documents for Online Learning. However KTM 6.X has many built-in Online Learning monitoring features and improved algorithms. These tools should now be used to mark documents for Online Learning instead of scripting. Implementing both will cause the unnecessary collection of documents and increase the time it takes to clean out the extraction set.

4. Increasing the maximum documents store for import: By default the maximum amount of documents that can be stored in the Online Learning directory is 2,000. This can be configured to up to 20,000 but is not recommended. Increasing the maximum storage will allow Online Learning

to be maintained less often, but will greatly increase the length of time it takes every time maintenance is done. It is recommended to keep the maximum storage to 2,000 samples.

5. Running multiple Knowledge Base Learning Services: Unlike KTM Server which can be configured to run in parallel on multiple servers, there should only be one server running the Knowledge Base Learning Service at a time. The Online Learning Server takes much less processing power than KTM Server, so there is no need to be running more than one service at a time.