Kofax Transformation Modules Database Locator

Application Note

Date	June 21, 2011
Applies To	Kofax Transformation Modules 3.5 and 4.0
Summary	This application note provides instructions and samples for the KTM Database Locator
Revision	1.0

Kofax Transformation Modules Database Locator Overview

The Kofax Transformation Modules Database Locator uses an associative index to match any record in a database with the document. This feature, along with its base product, is customizable per project basis, as demonstrated by information and samples provided in this documentation.

DATABASE LOCATOR

General

The Database Locator uses an associative index to match any record in a database with the document. The Database Locator also matches the document with one or more records in a database which are returned as alternatives. The best matching record is automatically selected as the field content. The Database Locator returns one index value for each column in the source database, which can in turn be assigned to an index field. The database has to be provided as a text file (such as a CSV file, which is comma separated values) first, and is imported once to generate the associative index that is used for the matching process. During runtime the associative index is updated automatically.

Database extraction can be defined multiple times for different sources and can manage several million entries (depending on the number of columns) with standard PC hardware.

For example, the Database Locator can be used to identify the sender of an invoice by using a supplier database. All that is required is enough OCR information to identify one or more possible records, which are returned as alternatives.

Another example is automatically extracting the ‘customer reference number’ by evaluating results returned by a Database Locator, which queries the customer’s database using name and address as a search term.

One final example is to use a Database Locator to match the document to the customer without the need for an explicit reference number.

These examples are some of the many solutions that the Database Locator feature can be used for

Performance

It is important to know that in an imported database's settings dialog, you can decide whether this database is to be loaded into memory or not. If you load it into memory, we can reach a lookup speed of less than 1 second on a database table with 10 columns and 1 million records. We suggest not loading databases into memory if they are larger than 5 million records. The best way is to test and see how much memory is consumed. If the database is not in memory, lookup speed is of course slower.

Confidence

How is the confidence for a possible match calculated? Let's assume all words on the document are spelled (and recognized by OCR) in the exact same way as in the database. If the database table has 10 columns (hence the records each have 10 fields) the confidence is calculated as:

$$\text{Confidence} = \frac{\text{Number of non-empty database fields found on document}}{\text{Number of fields in the record}}$$

For example, if the record has 8 fields filled with text (and 2 empty) and of these we find 6 on the document, the confidence is 6/8 = 75%.

If the record has 10 filled fields (none empty) and we find 6 of them on the document, the result is 6/10 = 60%.

If the record has only one field filled (That would be a bad record in a vendor database) and this field is matched, the result is 1/1 = 100%.

If you want records with empty fields to receive a penalty, here is a suggestion how to do that.

In the first case, where the result is 6/8, each of the 6 fields found on the document contributes 12.5% *6 = 75% to the record's total confidence. If one of them was only partially hit (OCR errors, difference in database and document content, etc), there is a penalty applied to this field's 12.5%.

Groups

It is important to make use of groups, of which there are 2 kinds: within a field and across fields. You can restrict each field's content to be found close together for example. If you have a NAME field in the record and it says "Jones Plumbing", you don't want "Jones" to be found on the top left of the document and "Plumbing" on the lower right. Instead you may want to set the field's inner grouping to "same line".

Grouping several fields is important if, for example there is a CITY and a ZIP field in the record. Here you want to add CITY and ZIP to Group1 and then apply "very close" to that group.

Importance

Make sure you apply the importance to each field in an appropriate way. In a customer or vendor database it usually makes sense to set the importance of the NAME field very high. Depending on overall document quality, you might even set it to "must exist" to avoid false positives. The CITY is usually not that important, because, in an invoice project, many vendors are in the same city.

Configuring a Database Locator

To configure a Database Locator, add it to the Extraction Design for the selected class.

Open test documents for this class before modifying the locator's properties. A Database Locator can then be configured by opening its properties.

To add a locator

1. In the Project Structure panel, select the class in which you want to add the locators.
2. Select Show Extraction Design ( ) from the Mode toolbar.
3. Click Add Locator ( ) from the Extraction Design toolbar.

4. Type a name.
5. Select the appropriate locator type from the Locator Method list.

To open test documents

1. In the bottom right panel of Project Builder, click Test Documents ( ).
2. Click Open Test Document ( ).
3. Click Browse.
4. Select a folder containing representative documents for a class and click OK.
5. On the Select Test Documents window, click OK.

To open a locator's properties

1. In the Project Structure panel, select the class containing the locator.
2. Select Show Extraction Design ( ) from the mode toolbar.
3. Select the locator.
4. Double click on the arrow symbol in blue. Alternatively, right-click on the locator and select Locator Properties from the context menu.

The properties for a Database Locator open in a new window which contains the following tabs: General, Fields, Search Masks, Regions, and Test Results.

General Tab - Database Locator Properties

Before you can modify these settings, ensure that the following tasks have been completed:

- The correct class has been selected.
- A Database Locator has been added.
- The relevant test documents have been opened in the Test Documents panel.
- The General tab of the Database Locator is used to select databases and to set the maximum number of alternatives, the confidence threshold and the penalty for empty fields in a record.

IMPORTANT: All databases must be added to the project via the Project Settings - Database tab. These databases are usually files containing delimited fields, such as CSV files. To add a database to the project, click on Database Settings.

Database

In order to use this locator, a database must be selected from the list of available databases.

The selected databases will be compared to OCR data during extraction. This means that if there is a record in a database that matches some content on a document, the record will be returned as an alternative. Depending on the similarities between a document and a particular database record, a confidence is applied to each alternative.

The first database you specify is the main database, which is mandatory for this locator. If fields on a document match a record in this database, they are returned as alternatives.

The second database is optional and is used to exclude document content from the list of alternatives.

The main database allows you to configure the physical relationship between fields on the Fields tab. The exclusion database cannot be configured in this way; so as a result, all fields must be located close together in order to be recognized.

Select existing database for locator

If a record matches the extracted data, it will be returned as an alternative. The amount of data that must match is configured on the Search Masks tab. This minimizes the effect of OCR errors in extracted data.

To select an existing database for locator, select the relevant database from the list.

Select a database with records which should not be found by this locator (e.g. own addresses).

The database contains records which are compared with the results of the main database. If there is a match, that record will not be returned in the list of alternatives.

NOTE: This database does not need to contain the same columns as the main database and the fields must be physically close together on the document. To select a database with records that should not be found by this locator, select the relevant database from the list.

Locator Algorithm Properties

The following properties can be configured to restrict which alternatives are returned.

Max Alternatives

This option allows for entering a number to limit the number of alternatives that will be returned by this locator. It is recommended that you have at least 2 alternatives available for comparison if the distance option is used.

To set maximum alternatives, enter a number in the field.

Min. Confidence

This option allows for specifying the threshold for the minimum confidence required for a match to be used an alternative. Only matches with a confidence greater than the threshold will be returned.

To set minimum confidence, adjust the Min. confidence slider to the desired amount.

Penalty for empty fields

This option allows for specifying the penalty for empty fields in the database. The maximum penalty is the percentage of empty fields in a record. The Penalty for empty fields value determines how much of the maximum penalty is applied.

For example, a database contains records with ten fields. If one record contains two empty fields, the maximum penalty would be 20%. If the Penalty for empty fields value is 100, the actual penalty applied to the field will be 20%. If however, the Penalty for empty fields value is 50, the actual penalty applied to the fields will be 10%.

To set the penalty for empty fields, adjust the Penalty for empty fields slider to the desired amount.

Now that you have selected databases and configured the algorithm properties, you can:

- -Configure field and group properties.
- -Configure search masks.
- -Configure regions.

- -Test the locator.

Fields Tab - Database Locator Properties

Before you can modify these settings, ensure that the following tasks have been completed:

- The required database has been added via the General tab.

  Once a database has been specified for this locator on the General tab, the two tables on this tab should automatically populate with the database field information.

- The physical relationship of database fields on a document can be specified on this tab to ensure that the correct information is extracted. This is done with a combination of field and group properties.

Field Properties

The Field properties list shows the database fields and the following additional settings.

Group Index

The Group Index can be used to indicate that field values should have some sort of identifiable geometric relationship with each other. For example, some items on a document, such as zip code and city name, are close together and can be grouped.

You can define groups by simply assigning a Group Index to one or more fields in the field list. The number of groups is limited by the number of fields.

If grouped fields on the document are not related as specified, the confidence of the record is decreased. Those fields that are assigned to a group are then listed in the Group properties area, where you can determine further settings for the group.

To set the Group Index

1. Select the field to be grouped.
2. Click on the value in the Group Index column to display a list.
3. Assign the Group Index to place this field in that group.
4. Repeat the above steps for each field to be grouped.

Distance

Use this property to specify the expected proximity of the words in a field. For example, a company name frequently consists of several single words on the same line, as in “Global Mortgages Inc.” In this case, you can set the distance property to same line by selecting it from the list.

To set the Distance

1. Select the field which needs to have its distance configured.
2. Click on the value in the Distance column to display the list.
3. Select the desired option from the list.

Importance

Select the level of Importance for the field: normal, low, very low, high, or must exist.

This setting indicates how critical the existence of the value is. For example, if a value for “Country” is of low Importance, there is minimal impact to the confidence of the record if it is not found on the document.

To set the Importance

1. Select the field that needs to have its importance configured.
2. Click on the value in the Importance column to display a list.
3. Select the desired option from the list.

Replace

If this option is selected, the results do not use the text actually found on the document.

Instead, the extracted text is replaced by the corresponding value from the database. Most of the time, you want the original record from the database to be used as the result rather than the sometimes incorrect data from the document. This option is selected by default.

To ensure a database field does not replace the OCR value

1. Select the field whose OCR data you do not wish to replace.
2. Clear the box to ensure the field is not replaced.

Group Properties

The Group Properties list shows the groups and the fields that are assigned to this group, if any. The choices for distance are very close, same line, near, and medium.

For each group you can define a Distance attribute. This is similar to the Distance setting for the fields, but it applies to the entire group. For example, if you group “Last Name” and “First Name”, the Distance for this group should be set to very close since the member fields are likely to be immediately adjacent to each other on the document.

To set the distance for a group

1. Select the group which needs its Distance configured.
2. Click on the value in the Distance column to display a list.
3. Select the desired option from the list.

Example

The Fields Tab - An Example

The content of the “Name” field appears on the same line in the document.

The “Property Address” field has all its words on the same line. The four address fields have been grouped together in “Group 1”.

The “Zip Code” field has a high priority. If this field is not present on a document, the confidence of the record is low.

“Group 1” contains the four address fields and they appear on the same line in the document. If the fields are not on the same line, the confidence of the record is decreased.

Now that you have configured field and group properties, you can:

- -Configure search masks
- -Configure regions
- -Test the locator

Search Masks Tab - Database Locator Properties

Before you can modify these settings, ensure that the following tasks have been completed:

- One or more databases have been added via the General tab.
- Search masks should be used when you do not expect all fields of a database record to be present on a document, but certain subsets of fields are sufficient to get a confident database match.

For example, an insurance company uses its customer database for several health insurance forms. None of the forms contain all database fields on a single document, yet a customer could be safely identified by one of the following combinations of database fields:

- First name, last name, insurance number
- First name, last name, date of birth
- First name, last name, street, zip code, city

A search mask is a subset of the available database fields. When a Database Locator is created, a default search mask which contains all fields in the database is created and the *All fields mandatory* option is not selected.

If any of the search masks are found, the alternative that is returned has a high confidence.

Mandatory Fields

If the *All fields mandatory* option is selected, then all fields in the search mask must be found on the document.

For example, you have a search mask that has first name, last name and insurance number selected along with the *All fields mandatory* option. All three must be present in order for a result to be returned.

If however, the *All fields mandatory* option was not selected, and only the first name and last name are found, the result has a lower confidence, such as 66%.

> **NOTE:** Fields that are not selected on the “Field Name” panel in the Fuzzy Database Options cannot be selected within the search mask and are disabled.

The Fuzzy Database Options can be found by selecting Project from the Project Builder menu options, and opening Project Settings. Within Project Settings is a Databases tab that contains the Fuzzy Databases being used. Select the appropriate Fuzzy Database by clicking on it once, and select Properties from the right navigation menu.

To add a search mask

1. Click Add.
2. Clear any fields to be excluded from the search mask.
3. Select *All fields mandatory* if required.

To delete a search mask

1. Select the search mask to be deleted.
2. Click Delete.

Now that you have selected databases and configured the algorithm properties, you can:

- Configure regions.
- Test the locator.

Regions Tab - Database Locator Properties

The Database Locator is one of the several locators that supports regions.

By default, locators operate on the entire page for every page in a document. To expedite processing, you can define regions that restrict the locator to portions of a page or to certain pages. This is useful if, for example, you know that an item always appears at the bottom of the first page.

To locate an item on any page in a document

1. Open the Properties window for the locator.
2. Select the Regions tab.
3. If it is not already selected, select the *All Pages* option from the *Enable Locator For* panel.

To insert a region for a locator method

1. Specify whether the locator is for *All Pages* or a specific page (such as *First page*, *Middle Pages*, or *Last Page*) using the *Enable Locator For* the area or the Page column for the locator.
2. Click Add to add a region to the list of regions.
3. Change the properties for the region (such as the Top or Left values) as desired.

To view the region on a document, select a document from Test Documents or Training Set. The document is displayed in the Document Viewer. The regions you add are displayed on the document. Instead of using Add, you can also draw the region directly on the document in the Document Viewer. You can also select and resize/reposition an existing region.

Manually Drawing a Region in the Document Viewer

To locate an item on the first page

1. Open the Properties window for the locator.
2. Select the Regions tab.
3. Clear the *All Pages* option if it is not already cleared.
4. Select the *First Page* option.

To identify an item in the lower 40% of the last page (for example locating a bank account number on an invoice)

1. Open the Properties window for the locator.
2. Select the Regions tab.
3. Clear the *All Pages* option if it is not already cleared.
4. Select the *Last Page* option.

5. Click Add.

   If a document is open in the Document Viewer, the region is displayed on the last page.

6. Set Top to 60%, Width to 100%, and Height to 40%.

> **NOTE:** You can also manually drag the region's boundaries to achieve the desired values.

Now that you have selected databases and configured the algorithm properties, you can:

- Test the Database Locator.
- If no databases have been specified, add them via the General tab.

Test Results Tab - Database Locator Properties

When Test is clicked at the bottom of the properties window, any matching database record is displayed along with its confidence. If the Document Viewer is open, all fields are highlighted.

For more information, see Testing Locators in the Configuration Guide for KTM 4.0.

Frequently Asked Questions/Informational Issues:

1. Why am I receiving incorrect return values when using a Database Locator?

There are many factors that may influence incorrect return values when using a Database Locator in KTM. The confidence for the Database Locator is calculated in the following manner:

- RecordConfidence = Sum(ColumnConfidence) / Number of Columns
- ColumnConfidence = Number of Matched Words / Number of words in Columns

The first equation will be extended by the significance of the single database columns if you select different weights for the columns (i.e., a factor with which you raise and lower the influence of a certain column):

- RecordConfidence = Sum(ColumnWeight * ColumnConfidence) / Sum(ColumnWeight)

It is easier to keep the first version in mind when thinking about the algorithm. For example:

- Record: "LCI GmbH"; "79199"; "Kirchzarten"; (3 columns)
  Matched words: "LCI"; 79199; Kirchzarten
  ColumnConfidence: 1 / 2 = 0.5; 1.0; 1.0
  RecordConfidence: 2.5 / 3.0 = 0.83

The results depend on the matched words. Sometimes, changing the delimiters in the database setup may affect the results received.

2. What may be the issue when I receive the following error message: "The database connection was established successfully, but the retrieval of the column information failed. Please make sure that the table exists and the spelling is correct: 'databaseo.sqlTableName'"?

When trying to set up the database connection to a SQL Server database using the DatabaseDialog object within KTM, you can leave out the 'databaseo.' from the table name.

For example:

Do NOT Use

Const cTableName = "databaseo.sql*TableName*"

Use

Const cTableName = "sql*TableName*"

3. What are some ways to improve Database Locator results?

Consider using a Database Evaluator.
From the Help Guide:

The Database Evaluator was designed to improve the results from the Advanced Zone Locator with the help of a Fuzzy Database. This is especially very helpful for hand printed forms where the quality of OCR is usually very low. Of course, this requires that there is a database available for the possible values printed on the form. The Database Evaluator compares a single field or a set of fields to values in a database. Depending on the actual settings for each field, the Database Evaluator can either replace all fields with the values from the database record, only replace a subset of the matched fields with database values or if there is no unique match with the database, it can return the original OCR result.

- Consider adjusting confidence thresholds.
- Consider the quality of documents that the environment may be using.
- Consider using search masks for fields that are having poor results.
- Test and modify the "additional delimiter" section
- Test and modify the "character to be ignored" section

NOTE: Please keep in mind that these are only suggestions, and not guaranteed ways to improve the search results of a Database Locator.

4. Where can I find an example of a Database Locator script/project?

There is a sample Database Locator project in the Examples folder of the KTM installation:

\Program Files\Kofax\Transformation\Examples

Please note that the above path will be the same for Ascent Capture 7.5 and for Kofax Capture 8.0. The file name is *DatabaseLookup.zip.*

This example project also contains Database Lookup examples as well.

5. What does the following error message represent: "<x> invalid records were not imported due to their column count being different from the column count of the first line (<y>)"?

The reason for receiving the error in question is a formatting issue with regards to a database file in the KTM Project. Whatever format the previous database file was in, the new entries or database must have the exact same format, or such an error can occur. This could be because of an added delimiter, space, etc.

If there are errors in the vendor file, then it will default to the cached version and should have the information from the previous database/vendor file.

The next recommended step is to review the updated database/vendor file for differences in formatting.

6. I have configured a Database Locator in KTM Project Builder. When I test the Database Locator, there are no results returned, though everything appears to be set up correctly.

What are some suggestions to help return results in the Database Locator?

- Check the database format via the *Database* tab in Project Settings of KTM.
  The field names listed in the properties of the Database options may not be properly delimited or may have an incorrect format. Work with the *Field Delimiter, Additional Delimiter,* and *Characters to Ignore* options to properly separate the database information. Data should be properly separated and organized such that relevant data is contained in the field columns.
- Check the Locator Algorithm Properties on the *General* tab of the Database Locator Properties.
  The *Min Confidence* may need to be reduced, depending upon the quality of the sample image.
- Check the *Search Masks* tab in the Properties of the Database Locator.
  Search Masks should be used when you do not expect all fields of a database record to be present on a document, but certain subsets of fields are sufficient to get a confident database match. Make sure that all fields that results are expected for are checked. If a field that has an expected result is checked, but no result is found, the Database Locator test will not return results.
- Check if Regions are being used in the Regions tab of the Properties of the Database Locator.
  If they are, temporarily remove them and test. There may be a sample document that has relevant data outside or inside of a region, and the settings for this tab may be incorrect. Change the settings to reflect the general coordinates of pertinent data.

7. How can I ignore spaces in the data on a document when using the database Locator?

For example, the database has a VAT number of 1234, but the document states 1 2 3 4. The data with the spaces is not found by the database Locator, but if the spaces are added to the database, then the data is found.

- OCR Substitution (this can **only** be used by a Format Locator)
- Use a script locator that ignores spaces and feed the output into the Database Locator via scripting.

8. When using the "Auto Update from Import File" feature, I understand that when a Batch is opened by KTM server, and the project is loaded, the database is refreshed if needed (assuming that 'Auto Update from Import File' is indeed selected). What .CRP file gets updated for the project? What exactly is a .CRP file in relation to KTM and the database being referenced?

A KTM project can be configured to "automatically update import file" when setting up a database. If this option is checked (set), the KTM Server will copy the newly updated txt file to the "pubtypes/custom/<*GUID*>/LCI.mailroom" directory when an updated database file is encountered:

- This update will occur when a Batch is processed by KTM Server.
- The current Batch Class does not need to be republished or re-synced with the KTM project to reflect the change in the database.

The .CRP files represent the binary version of the database files used in the project configuration. In addition, the .CRP file is used when loading the database into memory.

9. I want to extract Vendor information from 1 of 3 databases based on a Batch field that is set when the Batch is created in Batch Manager. I'd like for the database to be selected before validation. So when I get to the validation if the Batch field was "A" then I want the vendor information to have been extracted using database "A". If the Batch field was "B" then I want the vendor information to have been extracted using database "B". And if the Batch field was "C" then I want the vendor information to have been extracted using database "C". How can I achieve this?

Set up 3 different Doc Classes, subclassify them according to the value of the Batch field, then override the Database Locator in each subclass to use the database of your choice.

Example of subclassifying the documents:

```
Private Sub Document_AfterClassifyXDoc(ByRef pXDoc As CASCADELib.CscXDocument)
    Dim pXRootFolder As CASCADELib.CscXFolder
    Set pXRootFolder = GetRootFolder (pXDoc.ParentFolder)

    If pXRootFolder.Fields.ItemByName("CompanyNoFolderField").Text = "100" Then
      pXDoc.Reclassify("Class100")
    End If

    If pXRootFolder.Fields.ItemByName("CompanyNoFolderField").Text = "200" Then
      pXDoc.Reclassify("Class200")
    End If

    If pXRootFolder.Fields.ItemByName("CompanyNoFolderField").Text = "1400" Then
      pXDoc.Reclassify("Class1400")
    End If
End Sub
```

10. Can I change region position by script in a Database Locator?

Yes, this is possible in a script. Set up any region on the Database Locator, which will be modified by script.

Then, at runtime, the following code can dynamically change this region:

```
Dim odatabaseLocator As CscLocatorDef

Set odatabaseLocator =
Project.ClassByName("MyClass").Locators.ItemByName("databaseLocator")


Dim oRegion As CscLocatorRegion

Set oRegion = odatabaseLocator.LocatorRegions(0)


oRegion.Left = 10

oRegion.Width = 200

oRegion.Top = 20

oRegion.Height = 400

oRegion.Pages = 4
```

Run this code before the locator itself runs, and bear in mind that, if run in Project Builder, this will change the region as set in the project file itself (If this matters to you, cache the region boundaries before you change them, and set them back at the end of the script).

11. Is it possible to connect a Format Locator to a Database Locator? I want to tell the Database Locator what format of data it should look for.

The Database Locator uses a fuzzy search, so it should always find the correct entry if all data is present on the document.

For instance, if you want to look for the PO number:

KTM extracts all text from the document, where every word is being matched with the database. With the order number the Database Locator delivers a result with a high confidence which is the match.

Another option is scripting your own Database Locator that fetches the result from another Format Locator, delivers that result to the “Before Extract” event and, after that returns, the proper database entry.