What data/information does KTA capture and store for Online Learning?

3018862

Question / Problem:
Due to privacy of information concerns a customer may ask what data/information KTA captures and stores for Online Learning, as well as what control the customer has over the collection and storage of this information. The information in the next section should help to explain what data is collected and how it's stored.

Answer / Solution:

**Extraction Online Learning**

This type of online learning is used to improve extraction results for a project. Extraction online learning is restricted to a specific set of trainable locators. The Extraction Set is used to train a project for extraction during configuration. Online learning is enabled in the General tab of the Project Settings. Depending on how you plan to use online learning, additional settings can be found in the Advanced Online Learning Options window.

So, depending on which locators will be in use and which options are selected, this will determine what data is collected and how it's used. Data related to field coordinates, type of data, format information, etc. from successfully extracted and classified documents is used to build a set of documents for comparison to new documents in order to improve classification and extraction results. The following explanation is from the online documentation for Transformation Designer.

"**Kofax TotalAgility is designed to read semi-structured documents. Therefore, every project has a set of predefined fields for the most common items found on all types of invoices. These fields are almost always logically arranged on the document, and can be assigned to one of these locators.**"

**Each locator takes advantage of existing knowledge about the geometry of these items and uses that knowledge to improve data extraction.**"

So, generally speaking, Online Learning mostly collects non-live data information, such as page layout and coordinate information, however, depending on the types of locators used and the configuration of the project, there could be rare instances where live data might be associated with a locator property and thus collected during Online Learning. This might possibly occur, for instance, if an invoice number is associated with a specific vendor company. Below are some specific questions that are answered related to this collection/storage process that may also prove valuable.
Do we have control over which specific fields get captured? In other words, does the product allow us to be granular enough to make sure that a specific field (for example, a date of birth) never gets captured by Online Learning?

For specific Online learning the samples are marked (manually or automatically) for collection and manually downloaded in Transformation Designer into the new samples training set for the projects. This training is typically the responsibility of the system administrator who processes sample documents that are placed in a training set. The training session creates new extraction patterns that are stored with the project.

For Dynamic knowledge bases (DSKB) - Five types of knowledge bases exist which store field information. For example, if the date of birth field is not populated by one of the below locators, then online learning DSKB will not monitor/capture this field.

- Amount Group (*.kba) - stores extraction patterns for amount fields such as subtotal and total.
- Invoice Group (*.kbi) - stores patterns for invoice fields such as invoice number and invoice date.
- Order Group (*.kbo) - stores extraction patterns for order fields such order number and order date.
- Trainable Group (*.kbtgl) - stores patterns for arbitrary values in fields not covered by the other group locators.
- Trainable Table (*.kbtbl) - stores extraction patterns for various types of fields within a table.

Does Online Learning allow you to control for how long data is kept that is being captured? For example, if we are capturing a specific field like a Drivers License Number... can we make sure it gets purged after X number of months?

For specific Online learning, as per above, this training is typically the responsibility of the system administrator who processes sample documents that are placed in a training set.

For Dynamic knowledge bases, when they create a new generation and train it will overwrite the existing data. To perform the following steps to increment the “generation” number (related to subfolders called Gen1, Gen2, Gen3…), do the following.

1. in TD, Download all the new samples from the server (and deleting them from the server).
2. Import all the new samples from the new sample’s doc set into the extraction training set. [! Using the import button "Add to Training Set of Selected Class", not drag and drop]
3. Resolve any conflicts and re-train.
4. Release the project to production

Do we have full visibility of what live fields are being captured?
For specific online learning, as per above. For Dynamic knowledge bases, they are binary files which cannot be viewed or accessed.

Applies to:

<table>
<thead>
<tr>
<th>Product</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>KTA</td>
<td>All</td>
</tr>
</tbody>
</table>