ELO Smart Input

ELO Smart Input

Table of contents

ELO Smart Input

Introduction

Overview

ELO Smart Input is an intelligent assistant that helps you enter metadata for incoming documents semi-automatically.

This module analyzes the full text of a document and extracts data from it, such as addresses, sums, or e-mail addresses. You can then simply select from the suggested metadata options that are shown. The suggestions must be checked and saved manually. In the ELO Smart Input configuration, you can add regular expressions to search for specific information in documents and compare matches against existing database tables.

You can use ELO Smart Input with the ELO Java Client and the ELO Web Client.

How it works

ELO Smart Input is based on an incoming and target folder principle.

1. The incoming folder is where you save documents that have not yet been assigned metadata.
2. Use ELO Smart Input to enter the metadata for the documents.
3. After the metadata is entered, the documents are moved to a defined target folder.

The folder and appearance of the tile have to be defined in the *ELOsi Administration* interface first. You can then switch to the ELO Smart Input interface. You will reach the interface via the tile navigation in the ELO Web Client or ELO Java Client.

Use

Fig.: ELO Smart Input suggestions in the ELO Web Client

ELO Smart Input is an interactive assistant that helps you enter metadata for incoming documents semi-automatically. The module analyzes the full text of a document and extracts data from it, such as addresses, sums, or e-mail addresses. Users receive suggestions for automatically detected data, which they can apply to the metadata. Selected suggestions are highlighted in blue.

If the suggestion isn't correct, you can either select one of the alternative suggestions or complete the field manually.

Click *Save* to finish entering the metadata.

The document is moved to the filing path defined in the configuration.

What can be recognized?

Many patterns are offered out-of-the-box:

- E-mails, date values, times, IBANs, VAT IDs, URLs
- Addresses (Germany-wide)
- Sums (including gross value, net value, and VAT recognition)
- Values of ELO keyword lists

With a regular expression, you can define, extend, and make changes to custom patterns (such as a specific ID pattern). You can also import data sets from CSV or TSV files or databases.

Automatic classification

ELO Smart Input has an integrated automatic document classifier. For each document, the system identifies the likely document type (e.g. invoice, delivery note, advertisement) and what metadata form is most relevant. Of course, only metadata forms that have a mapping definition are selected.

The classifier is *universal* and self-learning, meaning it works for every language, document type, and metadata form. It learns with each document whose metadata has been saved via ELO Smart Input, becoming more and more accurate over time.

> **Please note**
>
> We recommend saving documents with the appropriate metadata form right from the start to quickly improve the classifier's learning curve.

If the document already has a metadata form a mapping has been defined for, the classifier is skipped. The system assumes that the document is already assigned the correct metadata form.

Differences between the ELO Java Client and ELO Web Client

Both clients have advantages and disadvantages when using ELO Smart Input.

The ELO Java Client features *Click OCR* to quickly complete the metadata with the click of a mouse. In the web view component, the date and list fields are rendered as normal fields.

The ELO Web Client has a date picker, lists with drop-down menus, and autocomplete, provided you use an up-to-date browser. Click OCR is not possible in the ELO Web Client.

Installation

Installation with the ELO Server Setup

This module is installed using the ELO Server Setup.

1. In the ELO Server Setup, switch to the *Applications* tab.
2. In your repository area, select the *ELO Smart Input* application from the drop-down menu and click *Add*.

Fig.: Adding ELO Smart Input in the ELO Server Setup

ELO Smart Input is added to the list of applications and installed after confirming on the next tab – *Confirm*.

Opening the 'ELOsi administration' interface

You can reach the ELO Smart Input administration interface in two ways:

-

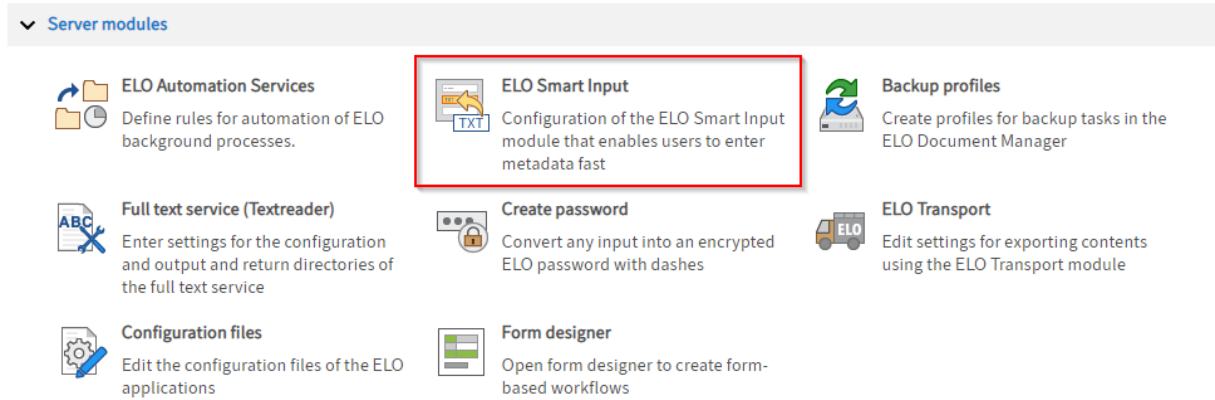Via *Server modules* in the ELO Administration Console

Fig.: Opening the administration interface via the ELO Administration Console

- Via the ELOis status page. You will find the *ELOsi Administration* button at the top right.
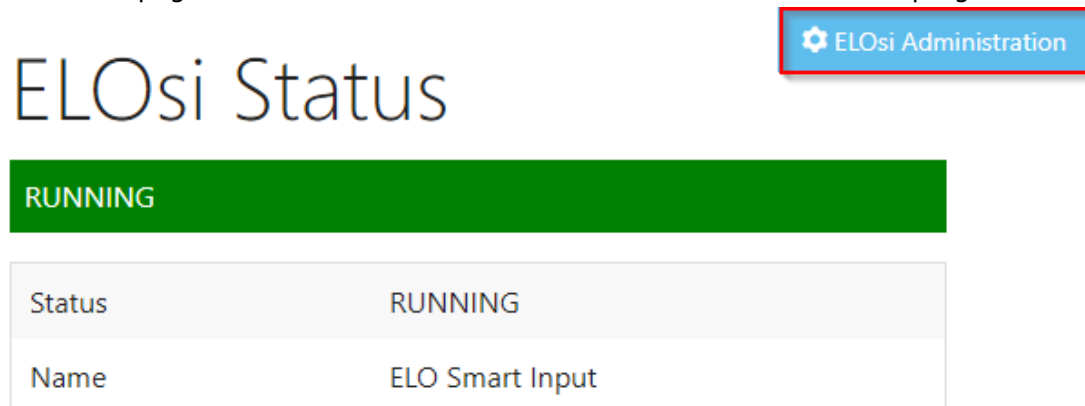

Fig.: ELOsi status, 'ELOsi Administration' button

Differences between version 20 and 21

The names of the content types have changed between these two versions. They should automatically be updated on start-up. If you encounter any problems, check to make sure the different mappings are correct.

The format of the file for the addresses has also changed compared with ELO 20. A *header* has been added. If you use your own files for addresses, update them and add them to the *header*. The new format also uses the TABULATOR key as a separator, but the separator ";" is also compatible, meaning that only the header should be added (with the TABULATOR key or ";" including the exclamation point).

Upgrading from version 12 to version 21

When installing ELO Smart Input using the ELO 21 Server Setup, a previous *ELO Smart Keywording* installation with ELO 12 is not recognized.

Follow the steps below before starting the ELO 21 Server Setup for this reason.

-

Uninstall ELO Smart Keywording:

- ○ On the Tomcat administration page, click *undeploy* next to *SK*.
- ○ If the former module was not uninstalled completely (may be the case with earlier versions):
    - ■ Stop Tomcat.
    - ■ In the folder *<elo-tomcat-dir>/webapps*, delete the file *sk.war* and the folder *sk*.
- ○ Delete the folder *<elo-tomcat-dir>/config/sk*.

- Install ELO Smart Input using the ELO Server Setup.
- Restore the old configuration:
    - ○ In ELO, rename the *Administration/ELOsk Base* folder *Administration/ELOsi Base*.
    - ○ If data tables were configured in ELO Smart Keywording, copy them to the new ELO Smart Input *data* directory.
    - ○ Restart ELO Smart Input.
    - ○ Delete the tiles in the ELO Smart Input administrative area and create them again.

Enable full text

For quick document processing, the relevant documents should already be filed to ELO and added to the full text database. For this to work, you will have to enable the *Add to full text database* option in the configuration for the pertinent metadata form in the ELO Administration Console.

Fig.: ELO Administration Console, 'Add to full text database' option

ELO Textreader has to be running for full text to work. Check the status in the browser:

```
<server name>:<Port>/<repository name>/tr?__cmd__=status
```

> **Information**
>
> You can also use ELO Smart Input without configuring this option. However, this would result in drastic delays, as OCR processing would be used instead.

Configuration

ELO Smart Input is configured for Germany by default. Some settings can be customized for use in other countries. These country-specific settings are made in the installation path and require you to restart ELO Smart Input.

Addresses

The SI package is supplied with German addresses as standard. You will find this CSV file in the folder */addresses/*. There, you can add additional files as address databases. Multiple files can be added. Make sure that these files are always UTF-8 encoded and that they have the same header.

> **Please note**
>
> The format of the file for addresses changed between version 20 and 21.

Street names are abbreviated (*Tübinger Str.*) or written out (*Tübinger Straße*). There is another setting in the `config.properties` for this:

```
# Default values
streetShort = str.
streetLong = straße
```

If you work with English addresses, for example, you should change it as follows:

```
streetShort = st.
streetLong = street
```

These settings are case-insensitive and only one abbreviation is possible.

To find out whether the street number is expected before the street name, which is common in some countries, you can use this setting:

```
# Default
streetAfterHouseNumber = false
```

Date format

Dates are recognized in various formats, including as text such as "1 January 2021". By default, German and English *written* date values are recognized.

If you want French and Italian date values to be recognized, adjust the `config.properties` as follows:

```
datesLanguages = fr, it
```

We recommend keeping the list of recognized languages to a minimum.

VAT

ELO Smart Input can recognize values such as the gross amount, net amount, and VAT provided they comply with the valid VAT rates. The German VAT rates are configured by default. For other languages, you can make changes via *<elo-si-config-path>/config.properties*.

```
# Valid taxes in %
taxes = 5, 7, 16, 19
```

> **Please note**
>
> In version 20, this variable was called `vatPercents` and has been renamed `taxes` in version 21.

If there is a VAT of zero percent in the country, you will also have to enter `0`.

Restricting options

If you want more or fewer results to be offered in the interface, you can use this setting:

```
# The maximum number of alternatives offered
choicesLimit = 4
```

Resetting the classifier

Sometimes metadata forms change, or you may want to file documents differently, for example because certain requirements have changed.

In such a situation, it may be beneficial to reset the classifier so that the configured form is no longer offered.

To reset the classifier, proceed as follows:

1. Stop ELO Smart Input.
2. Delete the raw data model files (document word statistics) under *<elo-sk-data>/classifier**.
3. Restart ELO Smart Input.

You can disable the classifier if you don't want it for some reason:

1. Edit the file *<elo-sk-config>/config.properties*.
2. Add `classify = false`.
3.

Restart ELO Smart Input.

Metadata mapping

Metadata mapping

From the standpoint of the software, a field in the metadata is simply a data type, such as *text*. ELO Smart Input first has to define what should be allocated to the field (e.g. e-mail, street name, IBAN).

This is done via *mapping*. In the bar on the left in the *Metadata mapping* area, you will see a list of available metadata forms. To define a mapping, select a metadata form.

Label	Field template	Content
sol.invoice.mask.vendorNo	IX_GRP_VENDOR_NO	tax_id
sol.invoice.mask.vendorName	IX_GRP_VENDOR_NAME	address.name
sol.invoice.mask.streetAddress	IX_GRP_VENDOR_ADDRESS_STREET	address.street_and_houseNumber
sol.invoice.mask.postCode	IX_GRP_VENDOR_ADDRESS_ZIPCODE	address.postCode
sol.invoice.mask.city	IX_GRP_VENDOR_ADDRESS_CITY	address.city
sol.invoice.mask.country	IX_GRP_VENDOR_ADDRESS_COUNTRY	address.country
sol.invoice.mask.iban	IX_GRP_VENDOR_IBAN	IBAN
sol.invoice.mask.VendorTaxId	IX_GRP_VENDOR_VAT_ID_NO	tax_id
sol.invoice.mask.invoiceNo	IX_GRP_INVOICE_NUMBER	number
sol.invoice.mask.invoiceDate	IX_GRP_INVOICE_DATE	date
sol.invoice.mask.dueDate	IX_GRP_INVOICE_DUE_DATE	date
sol.invoice.mask.invoiceNetAmount	IX_GRP_INVOICE_NET_AMOUNT	invoice.subtotal
sol.invoice.mask.invoiceTotalAmount	IX_GRP_INVOICE_TOTAL_AMOUNT	invoice.total
sol.invoice.mask.currencyCode	IX_GRP_INVOICE_CURRENCY_CODE	currency

Fig.: Metadata mapping

In the example above, the first four fields (fax, telephone, e-mail, IBAN) are individual fields. The last four fields (name, street, city, country) are group fields. The child entries of the group (in Fig. *Metadata mapping* the *Address* group) belong together. If you select a child entry, all relevant child entries of the group are filled in.

There are two group fields defined in ELO Smart Input: addresses and sums. Both group fields have a limited number of child fields.

In the *Filing path* field, you can define where documents with metadata are filed. For this path, you can use constants or variables from the metadata.

If no fields are defined for a metadata form, you can generate a mapping for all contained fields by clicking *Create all form fields*.

Alternatively, you can click *Add value* and *Add group of values* to create your own mappings.

In the *Name* area you will see the field name followed by the corresponding field template.

In the *Contents* area, select the relevant pattern. ELO Smart Input recognizes several predefined patterns, which you can select from the drop-down menu.

Preconfigured patterns

ELO Smart Input recognizes the following patterns by default:

- Addresses
    - Germany-wide
- Amounts
    - Everything with two decimal points
    - Gross, net, and VAT are recognized
- Date values
    - In different forms (06/21/2021 or June 21, 2021)
    - German and English formats as standard
- Currencies
    - €, $, £, ¥
- E-mail addresses
- IBAN
    - Global formats
- Telephone numbers
    - Heuristic
- VAT ID
    - European formats

You can also add your own regular expressions and data tables as described in the chapters Regular expressions and Data sets.

The following table contains example values for the content types (*patterns*) available in ELO Smart Input:

Content type	Example (recognized)	Example (output)
address.city	Stuttgart	Stuttgart
	stuttgart	Stuttgart
	STUTTGART	Stuttgart
address.countryCode	-	DE
address.houseNumber	(street name) 43	43
address.name	ELO Digital Office GmbH	*Heuristic*
address.postCode	70178	70178
address.state	-	*(Planned for a later release)
address.street	Tübinger Str.	Tübinger Str.
	TÜBINGER STR.	Tübinger Str.
	Tübinger Straße	Tübinger Str.
	Tübingerstr.	Tübinger Str.
amount	123456	*Must have two decimal points*
	123.456	*Must have two decimal points*
	1234.56	1234.56

Content type	Example (recognized)	Example (output)
	1234.56	123456.78
	1 234.56	123456.78
	1'234.56	123456.78
	1.234.56	123456.78
	1,234.56	123456.78
company	Example company	Example company
	Example company	Example company
credit_card	1234-5678-1234-5678	1234-5678-1234-5678
	1234 5678 1234 5678	1234 5678 1234 5678
	1234567812345678	1234567812345678
currency	€	€
	EUR	EUR
date	2021-03-07	2021-03-07 (saved in ELO as 20210307)
	7 March 2021	2021-03-07
	2021-03-07	2021-03-07
	7/3/21	2021-03-07
	March 7, 2021	2021-03-07
	30 Feb. 2021	*Invalid date*
email	info@elo.com	info@elo.com
IBAN	DE12345678901234567890	DE12345678901234567890
	DE 12 3456 7890 1234 5678 90	DE12345678901234567890
	BE68 5390 0754 7034	BE68539007547034
	SC18 SSCB 1101 0000 0000 0000 1497 USD	SC18SSCB11010000000000001497USD
invoice.total	105.00	105.00
invoice.subtotal	100.00	105.00
invoice.tax	5.00	5.00
number	12345	12345
person	Herr Max Mustermann	Max Mustermann
	Miss Jane Doe	Jane Doe
phone	+49 711 806089 - 0	+49 711 806089 - 0
phone	*Belgium:* 012 / 34 56 78	012 / 34 56 78
phone	*France:* 01-23-45-67-89	01-23-45-67-89
phone	*Poland:* (12) 345.67.89	(12) 345.67.89
phone	*United Kingdom:* 0123 456 7890	0123 456 7890
phone	*United States:* (123) 456 7890	(123) 456 7890
tax_id	DE123456789	DE123456789

Content type	Example (recognized)	Example (output)
	FR XX123456789	FRXX123456789
time	12:45:56	12:45:56
url	www.elo.com	www.elo.com
	https://supportweb.elo.com/de-de/elo-ecm-suite-21/module/elo-smart-input.html	https://supportweb.elo.com/de-de/elo-ecm-suite-21/module/elo-smart-input.html

Regular expressions

Basics

On the *Contents* tab, you can add custom regular expressions to enable ELO Smart Input to recognize additional patterns. Once the regular expression has been created, you can select it in the *Metadata mapping* area. Multiple patterns may have the same name. This allows you to use multiple different regular expressions for recognition such as `https?://.+` and `www\..+` as a URL pattern for a pattern name.

Fig.: 'Contents' tab: example regular expressions

In the *Name* area, enter a name for the pattern.

In the *Regular expression* area, you enter the regular expression.

If you only want to filter a portion of the defined regular expression, you can restrict it via the *Transformation* field. If you enter $1 as shown on the *'Contents' tab: example regular expressions* for example, only the content of the first bracket is taken over.

Clicking the *Test* button opens the *Tester* dialog box.
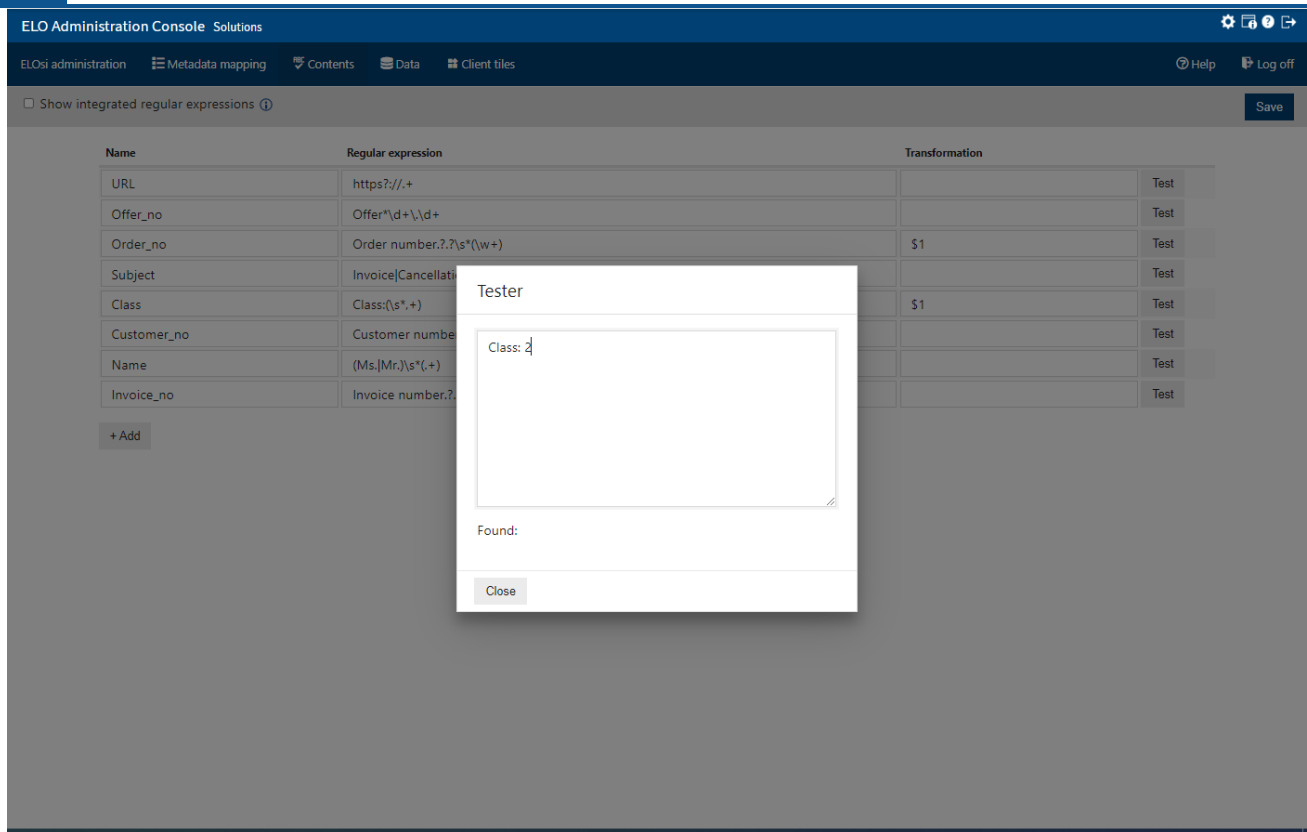
Fig.: 'Tester' dialog box for testing regular expressions

Here, you can test the defined regular expression. Enter an example text to the input field. The results of the regular expression search are shown below the input field.

Defining multiple regular expressions for a content type

It is also possible to configure multiple regular expressions per content type. We will illustrate this based on an example.

Scenario

The product IDs allow for an old format 123-456-678 and a new format ABC/123/XYZ.

Method

Instead of writing a complex regular expression to match both, you can define both formats separately.

Name	Regular expression
product_id	\d{3}-\d{3}-\d{3}
product_id	[A-Z]{3}/\d{3}/[A-Z]{3}

Everything matching the defined regular expressions is considered a valid product_id.

Extending preconfigured content types with regular expressions

Existing content types can be extended, see the following example.

Scenario

By default, the content type `credit_card` can only recognize *normal* credit card numbers such as 1234-1234-1234-1234. However, you want to recognize masked credit card numbers as well, such as 1234-1234-1234-XXXX.

Method

To enable this, you can add a regular expression also named `credit_card` such as:

Name	Regular expression
credit_card	\d{4}-\d{4}-\d{4}-XXXX

The content type `credit_card` would find two variants:

- 1234-1234-1234-1234 (default)
- 1234-1234-1234-XXXX (additional definition)

> **Please note**
>
> Group content types such as addresses or data sets cannot be extended with regular expressions.

Data sets

One typical scenario is searching for specific data in a document that cannot be identified by a regular expression, but instead by an external data source. Examples include project IDs, reference numbers, and customer names. With ELO Smart Input, you can integrate external data sources and use them to find IDs and add metadata.

Import CSV/TSV file

On the *Data* tab, you can add your own data sets for ELO Smart Input. Click *Upload file* to upload as many data sets as you'd like.

The file has to be a **.tsv* (tab-separated values) or **.csv* (comma-separated values) file. Note the following:

- The first line should be the header, which contains the column names (without whitespaces, without special characters).
- The file must be UTF-8 formatted.
- The first column is added as the index as default to be able to search in the document.

Example data set from generated random data:

```
id brand model year buzzwords
1C3BCBEB1DN774003 Toyota Corolla 2004 Switchable grid-enabled parallelism
WDCGG0EB9DG010196 Mitsubishi Pajero 1993 Grass-roots exuding support
3D4PG4FB1BT562959 Ford Explorer 1992 Networked multimedia approach
```

If the document contains 3D4PG4FB1BT562959 for example, it will be found. Values such as `Ford` or 1992 are also offered as additional information for the metadata.

Import via SQL database

You can also import data sets from an SQL database. In the *Data* area, click *Import from SQL database*.

There, you then have to enter the URL to the JDBC database, for example:

- `jdbc:sqlserver://<db-server>:1433;databaseName=testdb`
- `jdbc:postgresql://<db-server>:5432/testdb`
- `jdbc:db2://<db-server>:50000/testdb`
- `jdbc:oracle:thin:@<db-server>:1521/testdb.localdomain`

After clicking *Import from SQL database*, the *snapshot* is copied in full.

You can use any SQL statement. It is transferred directly to the database and the results are imported.

The first column is added as the index as default to be able to search in the document, independent of the database indexes.

Searching for multiple columns at once

Sometimes, you may need to search for another column or multiple columns at once in the document.

With CSV/TSV files, this can be done by placing an exclamation point before the column name.

Example

You want to know what car the document is about. The brand, model, and year of manufacture should be recognized in the document to identify the car. The file would then look as follows:

```
id !brand !model !year buzzwords
1C3BCBEB1DN774003 Toyota Corolla 2004 Switchable grid-enabled parallelism
WDCGG0EB9DG010196 Mitsubishi Pajero 1993 grass-roots exuding support
3D4PG4FB1BT562959 Ford Explorer 1992 Networked multimedia approach
```

You can then identify the associated ID or keywords.

You can do the same thing in *SQL imports*. There, you have to enter the indexes in the last field.

In the example recognizing the car, this would look as follows, without an exclamation point:

```
brand, model, year
```

Use data set

Once you have added a data set to ELO Smart Input, you can select it on the *Metadata mapping* tab and add it to a metadata form. There is no need restart.

Tiles

Client tiles (adding tiles)

On the *Client tiles* tab, you can generate web views for the tile navigation in the ELO Java Client and ELO Web Client.

Fig.: 'Client tiles' tab

ID: Enter an ID of your choice.

Input directory: Enter the GUID or path of the folder you want to monitor.

Name: Enter a name for the file.

Icon: Enter a name for the icon.

Add: Click *Add* to create the tile.

After restarting the ELO Java Client or clicking *Refresh* in the ELO Web Client, the tile appears in the tile navigation.

The purpose of the tiles is to be able to use ELO Smart Input for different folders with different target groups and rights.

FAQ

Can ELO Smart Input replace ELO DocXtractor?

ELO Smart Input and ELO DocXtractor are both products that analyze electronic documents while making it easier for users to capture information. However, they offer different benefits and functionality.

ELO Smart Input is a metadata assistant that helps users to import document information from metadata.

ELO DocXtractor is a document analysis system with advanced automation capabilities for knowledge-based extraction of data from any document.

The following table will help you decide which product suits your application.

|	ELO Smart Input	ELO DocXtractor
Functionality	Interactive metadata assistant	Fully automated data entry
Classification		
Can be configured	□	□
Self-learning	□	□
Text-based	□	□
Layout-based	□	□
OCR text recognition		
Digital documents	□	□
Scanned documents	□	□
Multi OCR engines	□	□
OCR fine-tuning	□	□
Recognition		
Common patterns	□	□
Text-based	□	□
Position-based	□	□
Custom regular expressions	□	□
Lists	□	□
Data sets	□	□
Barcodes	□	□
Tables	□	□
Scripting	□	□
Checks	□	□

You will find a detailed comparison of the two products in the ELO Community news area.

Is there an option to perform automatic/regular updates?

Not currently. This is planned for a future version, however.

Are there any size limits?

No. The search is optimized and should also work fine with large volumes of data.

Troubleshooting

ELO Smart Input is slow

Check whether the documents processed by ELO Smart Input have been added to the full text. If the documents are not in the full text, the system falls back to OCR.

Client error: HTTP Status 400 – bad request

The ELO Indexserver proxy setting contains a faulty URL or the ELO Indexserver was not restarted.

Client error: HTTP Status 401 - unauthorized

To show this web app in the ELO Java Client and ELO Web Client, you have to configure forwarding via a proxy URL. Check whether the URL in the *proxyURL* line leads to the status page. If this isn't the case, check the proxy configuration and restart the ELO Indexserver.

ELOsi Status

❖ ELOsi Administration

RUNNING	
Status	RUNNING
Name	ELO Smart Input
Version	20.00.010
BuildDate	2020-04-20T16:09:35Z
webappconfigdir	E:\\ELO\\DE\\config\\si-EXTEN\\ELO-DE
proxyUrl	https://doksrvint01vm:8083/ix-EXTEN/plugin/de.elo.ix.plugin.proxy/si/
ixUrl	https://doksrvint01vm:8083/ix-EXTEN/ix
ixUsername	ELO Service
ixPassword	******
dataPath	E:\\ELO\\DE\\data\\si-EXTEN\\ELO-DE
vats.1	1.07
vats.2	1.16
vats.3	1.19
choicesLimit	4

Fig.: ELOsi status, 'proxyURL'

Client error: HTTP Status 404 – not found

The ELO Indexserver proxy setting is missing or the ELO Indexserver was not restarted.

Client error: HTTP Status 503 – service unavailable

On the Tomcat status page, check whether the *si* web app is running.

1. Open the *ELO Application Server* overview page in your browser: `<server name>:<port>/manager`.

2. Select the web app *si*.

3. The ELOsi status page opens: `<server name>:<port>/si-<repository name>/status`.

The *Status* should be RUNNING.