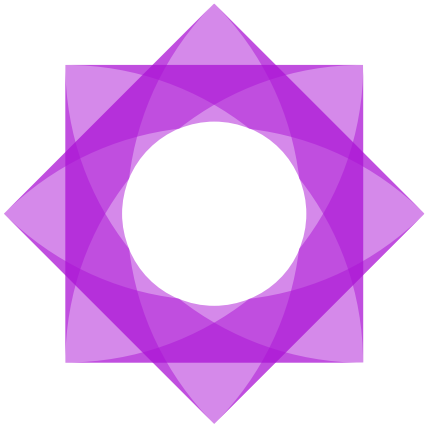


Lasernet 10.

Lasernet OCR 10

Adam McStravick, Torben Pedersen
Revision 5
2023-08-15

Formpipe.

Formpipe.

Contents.

1 Introduction.

1.1 Who Should Use This Guide?

This guide is written for end users who are responsible for the incoming document workflow. It is intended primarily for reference purposes and covers the various functions in Lasernet OCR Editor.

2 Terms of Use.

3 Lasernet OCR Editor and OCR Engine.

3.1 Introduction

Lasernet OCR Editor and Lasernet OCR Engine are a client / server solution used for maintaining incoming business information and the easy extraction of data for import directly into your workflow system. By setting up simple text recognition, you can work with data fields in forms such as invoices, order confirmations etc, from any of your business partners.

The only requirement is that documents must be delivered to a channel available to the Lasernet Server, as one of the following file types:

- PDF
- Text file
- TIFF

OCR Editor is included in two of the Lasernet Input Management packages, which support the following incoming formats:

Package	Input Format	Licensed modules in package
Input Management Basic	PDF, PDF/A with embedded text data	PDF to Text Excel to XML**
Input Management Standard	PDF, PDF/A TIFF with embedded text data or images*	PDF to Text Excel to XML** OCR Module** Tesseract OCR***
Input Management Full	PDF, PDF/A, TIFF with embedded text data or images*	PDF to Text Excel to XML** OCR Module** Tesseract OCR*** XML Out

The Barcode Reader module is required to extract barcodes into readable characters.

* For converting images to text, Tesseract OCR module is required for OCR scanning of TIFF files.

** An additional license for XML Input (Form Engine) is required.

*** Hand-written text, right-to-left and Asian languages are not supported. The module is trained and bundled with the following language packages: Danish, Dutch, German, English, Finnish, French, Icelandic, Italian, Norwegian, Russian, Spanish and Swedish.

Excel to XML conversion is also included in the Input Management packages. This process is not covered in this documentation because XML is not a supported input format for the OCR Editor application.

More information about how to configure the Tesseract OCR module, for optical character recognition,- is available in the Lasernet Developer manual.

3.2 Installers

The Lasernet system administrator installs and configures the back-end for Lasernet Input Management that includes the following list of Lasernet components:

Server installation

1. Lasernet Server installer
    - Lasernet Service
    - Lasernet Configuration Server
    - Lasernet Monitor
    - Lasernet Server License Manager
2. Lasernet Web Client installer
    - Laseret Web Server Service

Client installation

1. Lasernet Developer installer
    - Lasernet Developer
    - Lasernet Developer License Manager
    - Lasernet Monitor
2. Lasernet Client installer
    - Lasernet Client
3. Lasernet OCR installer
    - Lasernet OCR Editor

ℹ️ For more information on how to install and configure Lasernet applications, please refer to the Lasernet Installation Guide.

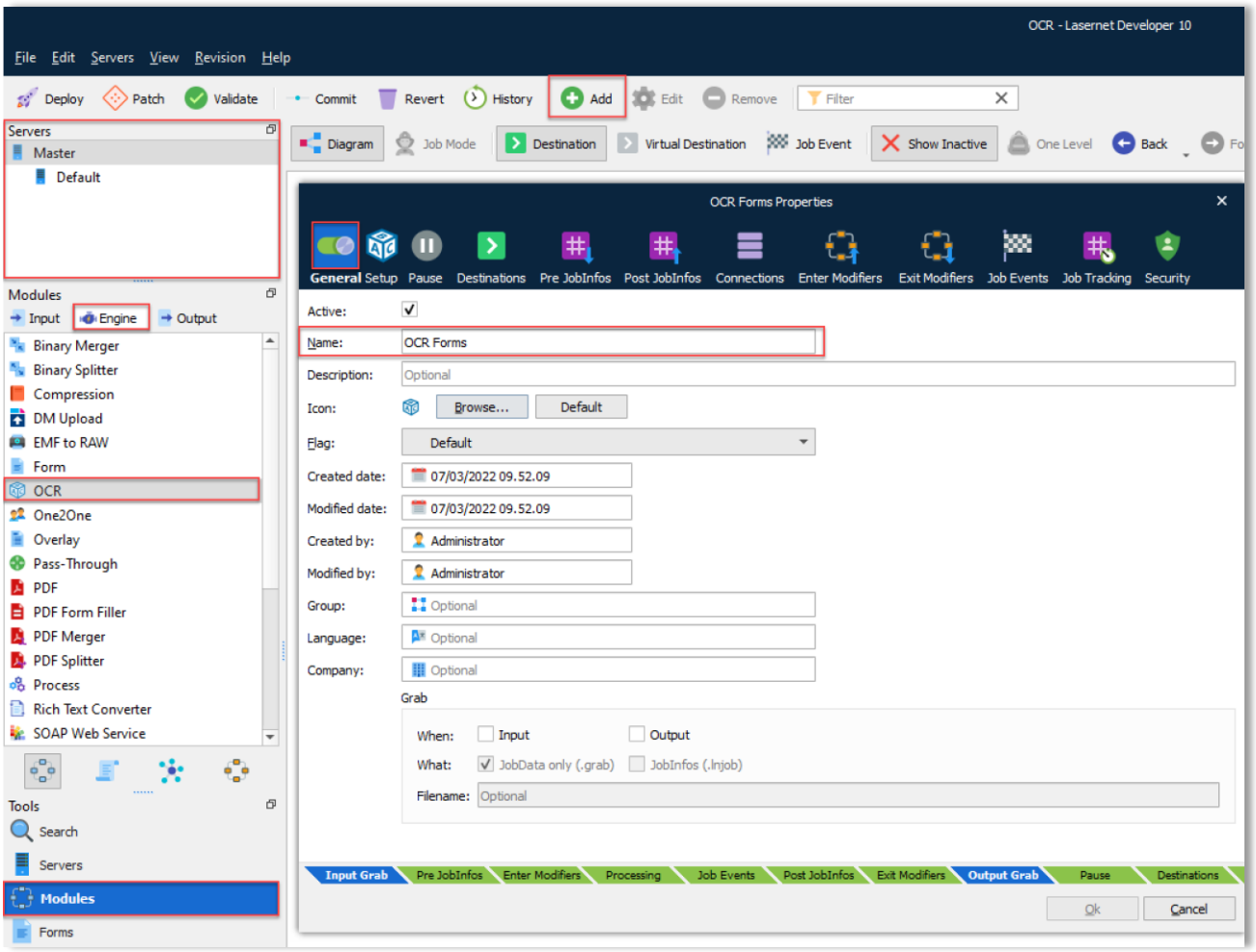
3.3 Getting Started – Lasernet Developer

3.3.1 Configuring OCR Engine

This chapter is intended for advanced users who configure the back-end for:

- OCR Engine
- OCR Fields
- Validation
- Dictionary
- Workflow

From the Lasernet Developer, you must select the server instance from which you want to run the OCR Engine. Select Modules ► Engine ► OCR and click the Add tool to create a new OCR Engine for your input management workflow.



Type any name for the new OCR Engine. Select the Setup tab to define the list of documents for which you want to set up OCR scanning, and which fields you want to extract from the documents. The information

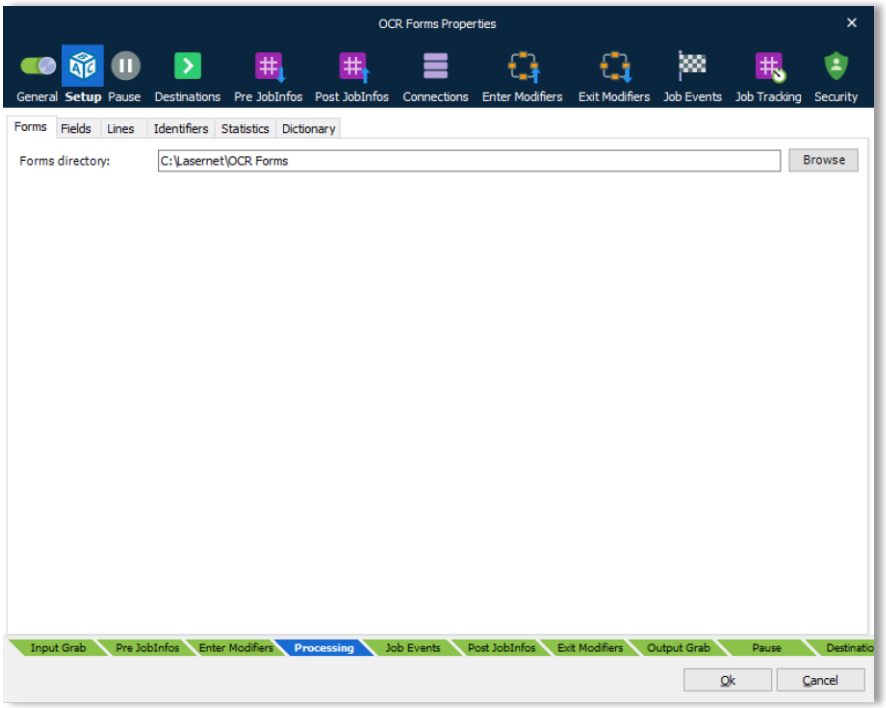
defined in the OCR Engine is used by the OCR Editor to help the end user to set up documents, and by the OCR Engine to automatically extract the data via the Lasernet Server.

3.3.2 Forms

In the Forms tab you define the forms directory where OCR Engine will store:

- OCR Forms with the grabbed PDF
- OCR Forms, created in Auto Capture mode, with the captured PDF

Note: We strongly recommend that you create a backup of this folder.



Forms directory — Folder path for the OCR Engine to store the list of OCR forms, grabbed PDF and auto captured forms

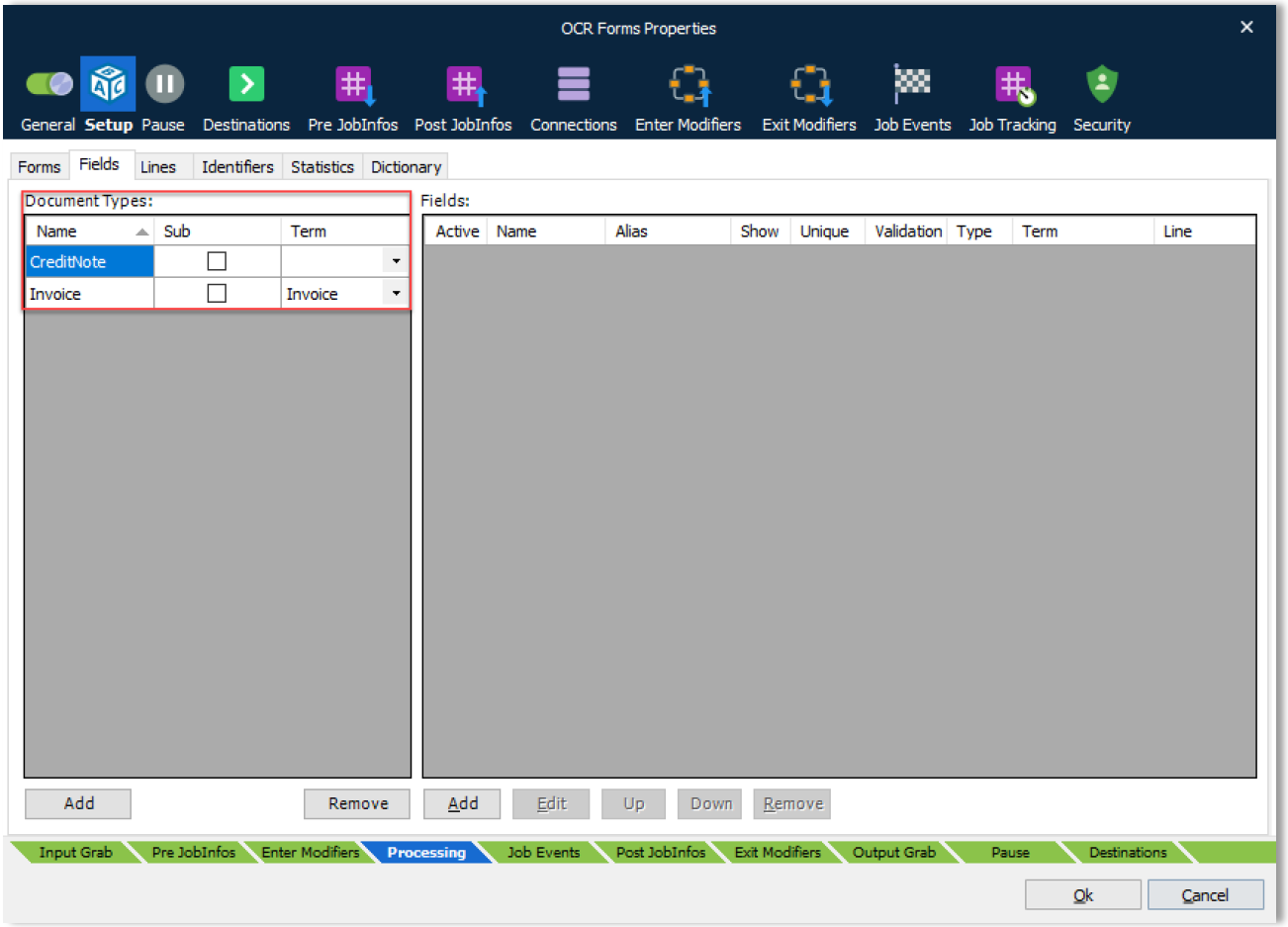
⚠ **Caution:** Do not use the same directory as the “Server runtime data directory”, defined in the server settings of your Lasernet Server, or you will corrupt your OCR forms.

3.3.3 Fields

In the Fields tab you can add the Document Types, Field Names and Data Types required by a third-party workflow system. You can customize your OCR Fields with aliases to be used by OCR Editor, rights to show an OCR Field in the Lasernet Client and with validation rules to fail a job.

If the Dictionary is activated, you can create a link between documents and field types added to the OCR Engine, and terms added to the Dictionary. This will optimize the process of creating OCR Forms in the OCR Editor, or AutoCapture data in documents to automate the OCR process as much as possible.

Document Types

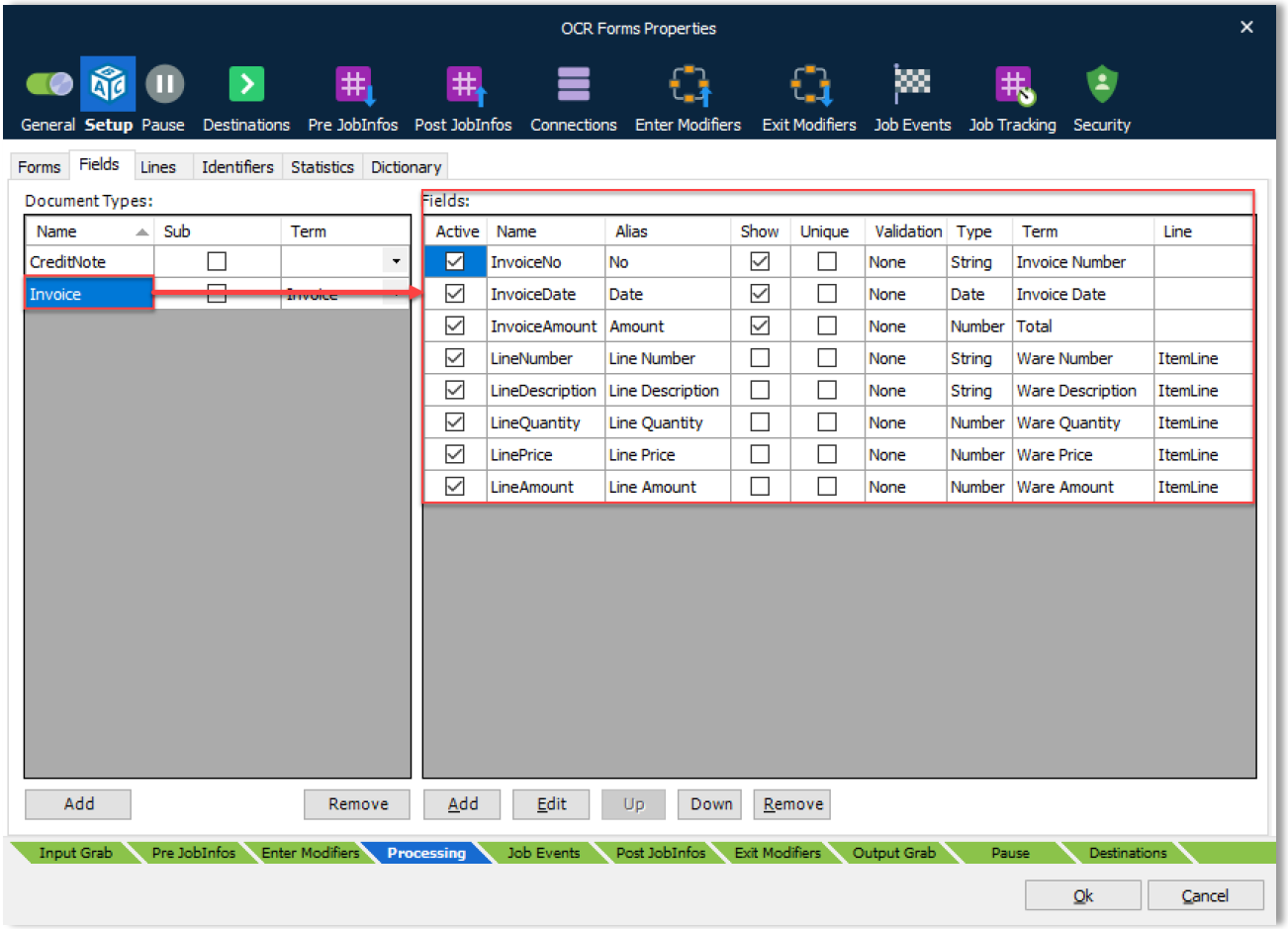


Add a **Name** for each of the document types from which you want to extract OCR data. For example, common document types are invoices, credit notes, orders etc. The document types must reflect the various documents from which you want to extract data via the OCR Engine. For each document type, you must add a list of fields that reflect the requirements of your external workflow system to which you want to export data.

Select the **Sub** checkbox when there are minor variants for specific documents types. For example, an invoice vs. proforma invoice, where you mark a document as a sub-type and allow the end user to select both the main document type and the sub-type when creating an OCR Form in the OCR Editor. This can help you improve the maintenance of OCR Fields by sharing field types between variants of the same document type.

Term The Term column is a binding between the OCR Field added to the OCR Engine, and a Term stored in the Dictionary database for automatic recognition of documents. Values are only accessible from the drop-down menu if the Dictionary Server is activated and is used by the AutoCapture feature, if activated.

Fields



Field Names – Any Field Name added to a document will be listed and accessible from the OCR Editor.

The Name column must contain the field name to be included in XML data created as output from the OCR Engine. If an illegal tag name is entered you will prompted with an ! icon.

The Alias column must contain the logical field name seen by the end user in the OCR Editor. Logical names will help the end user to locate the required data fields in the incoming OCR documents.

Select the **Show** checkbox to enable viewing of the field name, or alias name (if entered) in the Lasernet Client. Note that if this field is not activated, validation cannot be performed by the end user in the Lasernet Client.

The Validation column will contain the value “None” if no validation for this field is required, or “Automatic” if a validation rule has been entered. Select a Field Name and click **Edit** to set up validation rules.

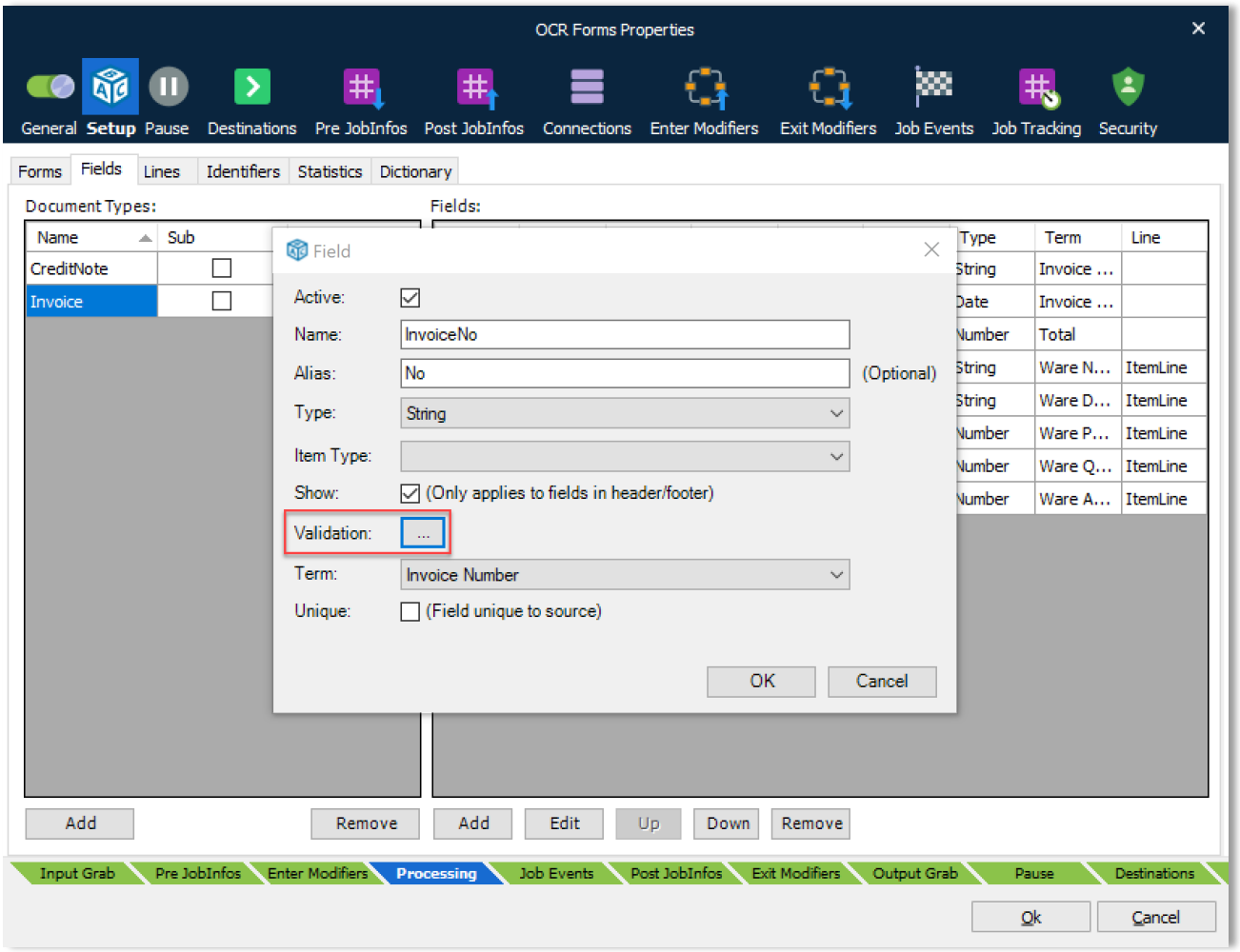
Each field can be defined as a Type, such as string, number or date. Use number and date formats for converting various incoming number and date formats, via Regional Profiles, into a standard format. Numbers will be converted to 99999.9999 (four decimals) and dates formatted to yyyyMMdd.

The Term column is used to connect a field name with the name of a term in the dictionary. A term consists of one or more aliases that appear in job data. When the dictionary understands a phrase, it will automatically be connected to the defined field name.

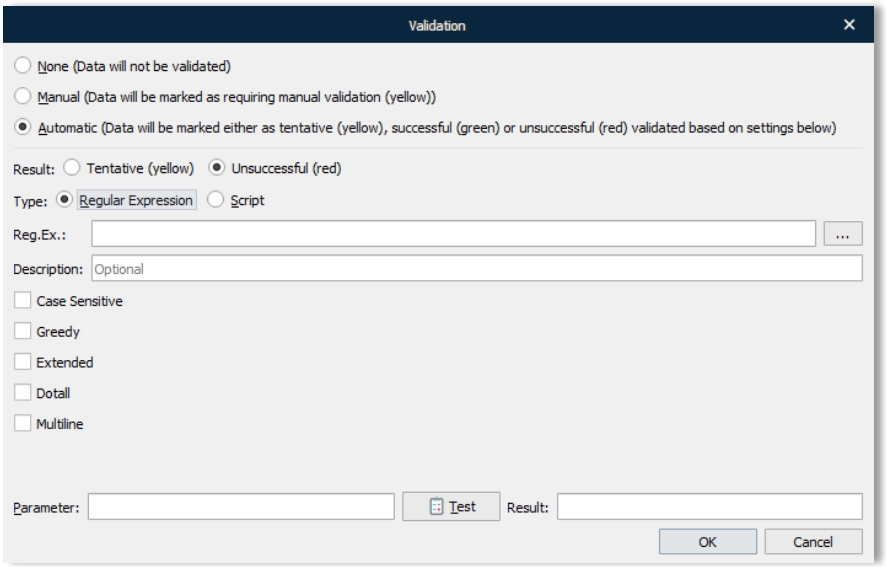
The Line column can associate an OCR Field to a specific type of item line added in the Lines tab. Binding OCR Fields to specific item lines will improve the steps for adding OCR Fields to an OCR Form, via the wizard in the OCR Editor.

3.3.4 Validation

Validation is used to validate the contents of an OCR Field. Metadata that does not match specified conditions will be marked either as tentative (yellow) or unsuccessful (red) in the Lasernet Client, which is used for approval of documents in the workflow.

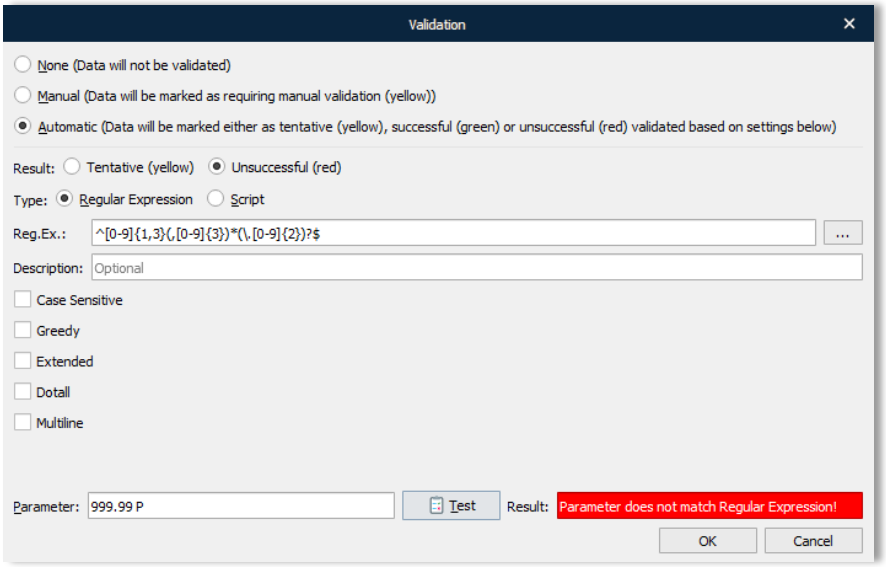


Select a **Field Name**, click **Edit** and click the Validation **...** button to edit or create new validation rules.



Select **Regular Expression** or **Script** to validate the contents of an OCR Field.

Validate with a regular expression

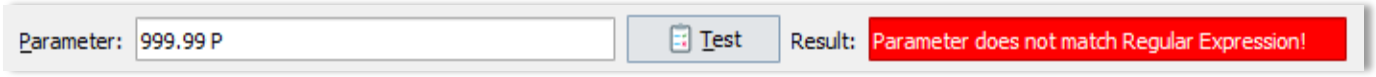


In the example we have created a regular expression matching a US formatted number, using a comma as the thousand delimiter and a dot as the comma delimiter.

Regular Expression: ^[0-9]{1,3}(,[0-9]{3})*(\.[0-9]{2})?$

The $ sign at the end of the regular expression is essential to match the full string and not a part of the string.

The **Parameter** field and **Test** button are useful to validate the expected result of the regular expression. This functionality is for online testing only in the Lasernet Developer.



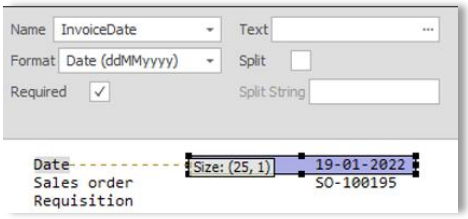
For more information regarding regular expression syntax, we recommend using the internet for reference.

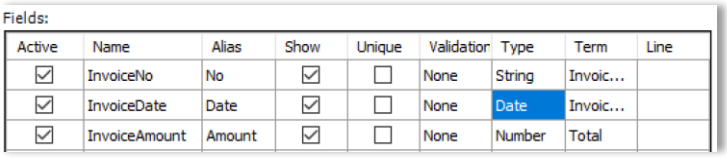
Validating dates and numbers

All fields defined with date or number as the field type must be validated against the internal date or number format handled by the OCR Engine; not as the format present in the document.

The internal format for dates is: yyyyMMdd

The internal format for numbers is: 99999.9999 (dot as the decimal splitter and four decimal places)

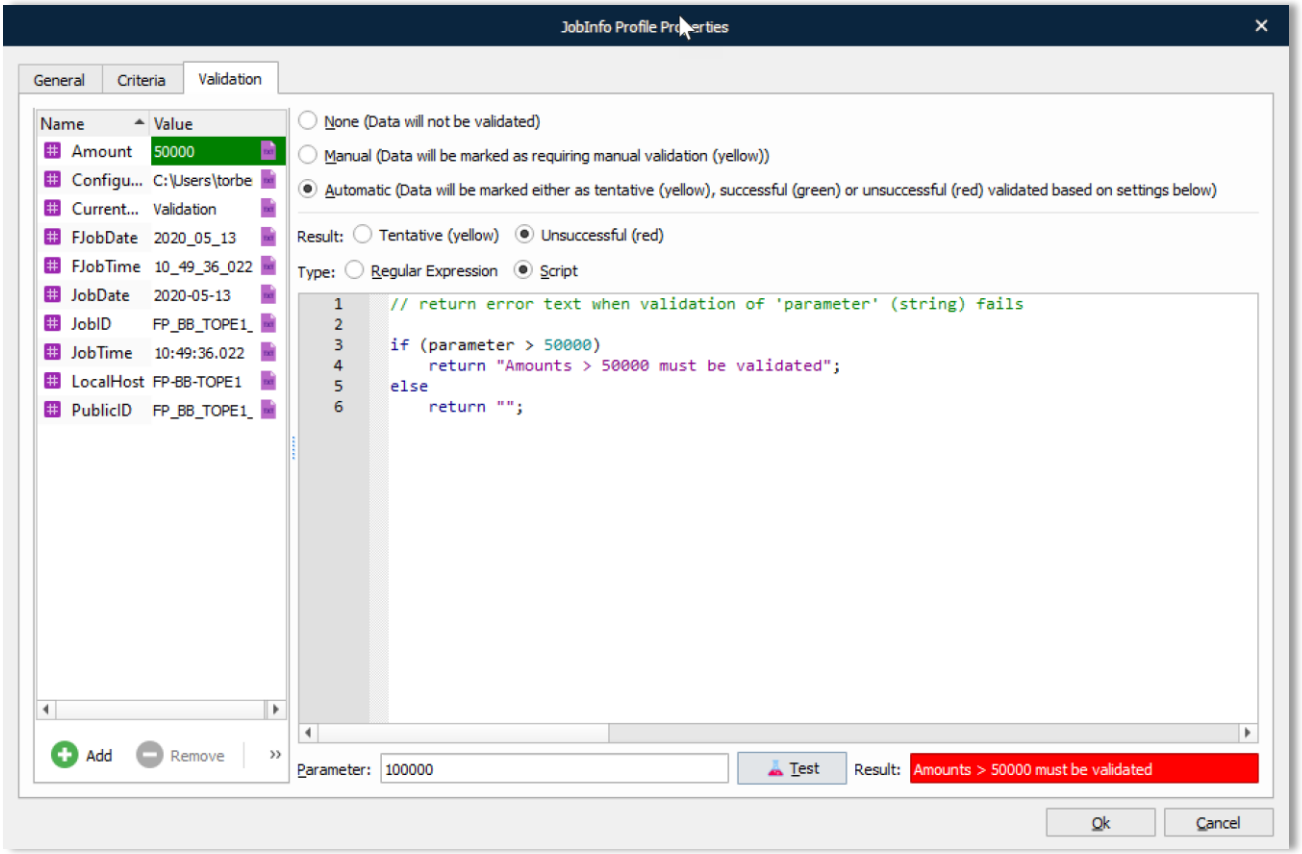




In this example, the date 19-01-2022 is extracted and converted into the internal format 20220119.

For the date field you must define a regular expression that validates to the internal format of yyyyMMdd. For numbers you must validate to the format 99999.9999

Validate via JavaScript



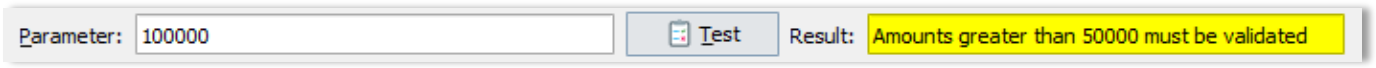
In this example we have created a JavaScript matching if a parameter is higher than a specific amount. The **parameter string** is essential and contains the contents of the JobInfo to be validated. The expression will return an empty string when validation of the parameter is successful and error text when it fails validation.

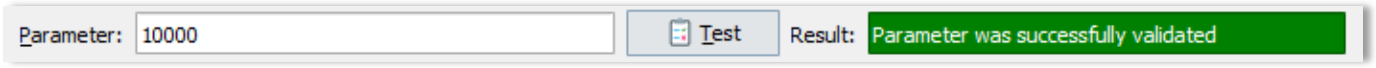
```
if (parameter > 50000)
  return "Amounts > 50.000 must be validated";
else
  return "";
```

To validate against a **JobInfo** included in the script, click **Add** to add the JobInfo **Name** and **Value** in the JobInfo grab window.

```
if (parseFloat(parameter) > parseFloat(job.getJobInfo("Amount")))
  return "Amounts > 50.000 must be validated";
else
  return "";
```

The **Parameter** field and **Test** button are useful to validate the expected result of the JavaScript. This functionality is for online testing only in the Lasernet Developer.





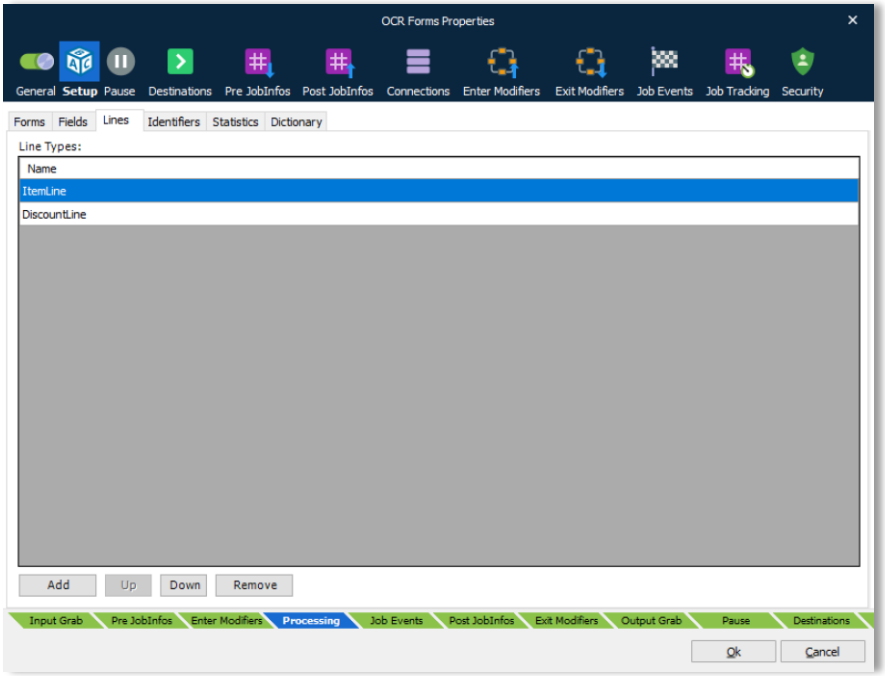
Supported JavaScript functions

The JavaScript tool in the OCR validation solution has support for a limited range of system functions compared to the Job Engine. The supported convert functions, job classes and arrays are:

- Static functions:
  - Convert.toDate
  - Convert.toDateRP
  - Convert.toNumber
  - Convert.toNumberRP
- Job class:
  - All functions in the job class (see the Scripting manual section 4.2).
- Arrays:
  - Azure commands
  - Database commands
  - Modifier commands
  - SharePoint commands

Note: The system functions are not supported for validation in Lasernet Meta, where validation runs on client side and not server side.

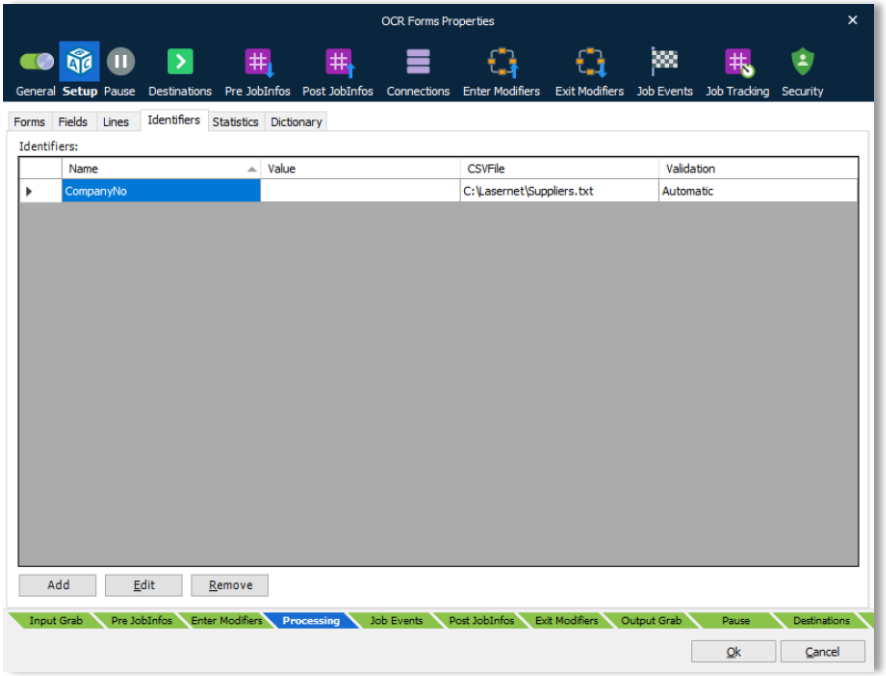
3.3.5 Line Types



Line Types added to Line tab are assigned to the last Document Type selected in the Fields tab. When the list of Line Types is added, you can bind them to specific OCR Fields via the Line column in the Fields tab.

The Line Types with OCR Fields will be embedded in the output job if recognized by the OCR Engine. The list can be left empty if no lines are captured from the document. When recognizing Line Types in the body area of the incoming documents, the order of Line Types is important – Line Types at the top of the list will have the highest priority when the OCR Engine is recognizing lines. Move Line Types up and down to set the priority.

3.3.6 Identifiers



Identifiers are used for inserting fixed field names and values in the output data in OCR Forms from the Lasernet Client or OCR Editor. Set a default **value** for the field if required or leave empty.

Note: Only characters from A–Z are recommended as valid Identifier names. Spaces are not allowed.

Set the path and filename to a local CSV file stored on the Lasernet Server. CSV files most be stored in the following format:

```
“Header1”,“Header2,“Header3”
“Value1,”Value2”,”Value3”
“Value1,”Value2”,”Value3”
```

The list of records added to the CSV file are included in the drop-down menu accessible from the Lasernet Client and OCR Editor. The value of the first column can be selected and used as the value for the field. Other columns are for viewing purposes only.

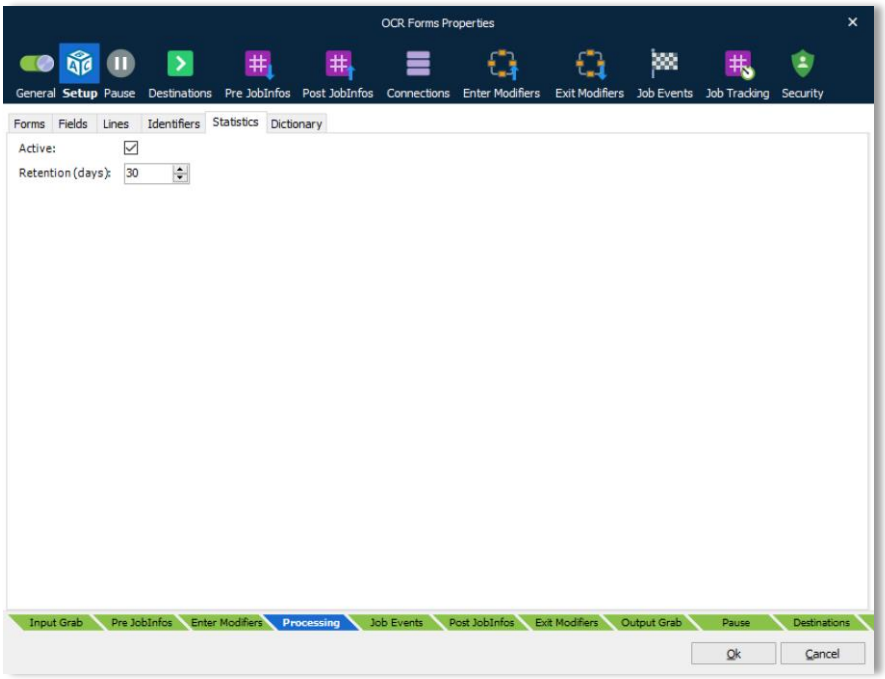
- Header names (Required)
- Field separator (Comma only)
- Quotes: (Not required, but recommended)

Note: A maximum of 6 field columns are recommended for viewing purposes.

3.3.7 Statistics

Activate statistics and save information about the capture rate of OCR Fields for each document processed by the OCR Engine. This is a useful tool if you want to prioritize and improve the quality of extracted data for specific OCR Forms, particularly where there is a high load of incoming documents.

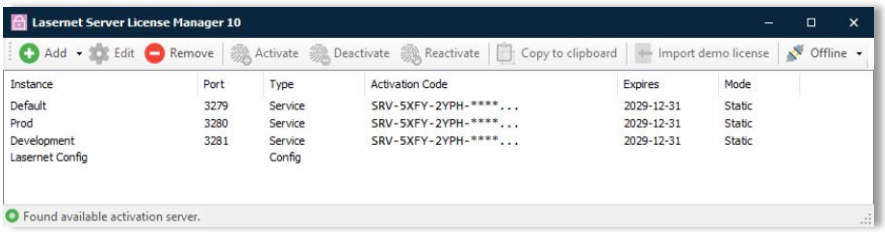
The statistics are accessible for the users in the OCR Editor and do not contain any confidential information. The statics contain information about number of hits per document and an average percentage of the number of captured OCR Fields for each OCR form.



3.3.8 Dictionary Service

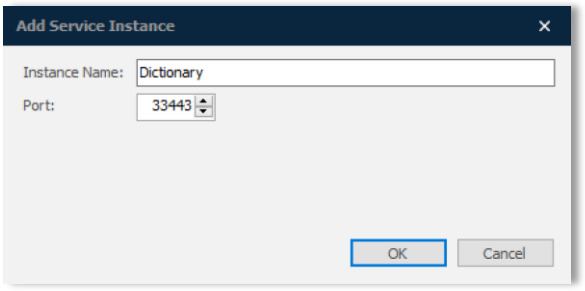
The Lasernet Dictionary Service is an optional feature, which runs with a Microsoft SQL Server LocalDB database as the back-end, to automate the capturing of OCR data in documents.

The Lasernet Dictionary must be installed and configured with a service name and a port number in the Lasernet Server License Manager.



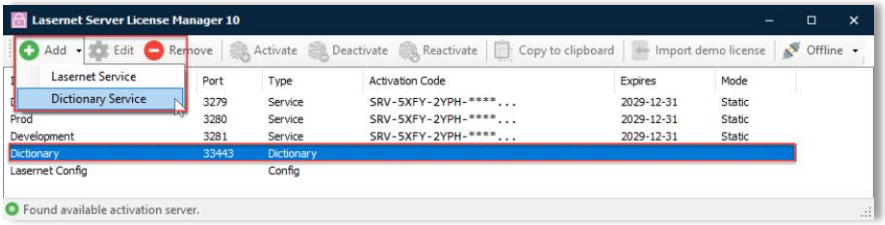
We recommend that you configure the same port number as for the Lasernet Config Server, where the default port number is 33443.

Click **Add ► Dictionary Service** then enter an Instance Name and a Port number.

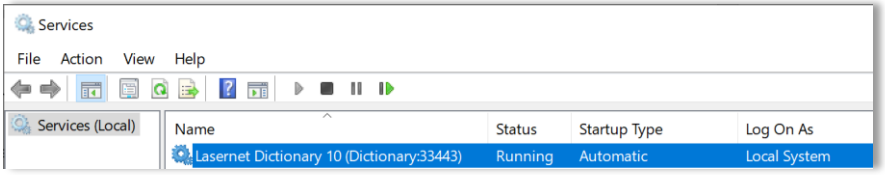


A new Dictionary will be created with the selected Instance Name.

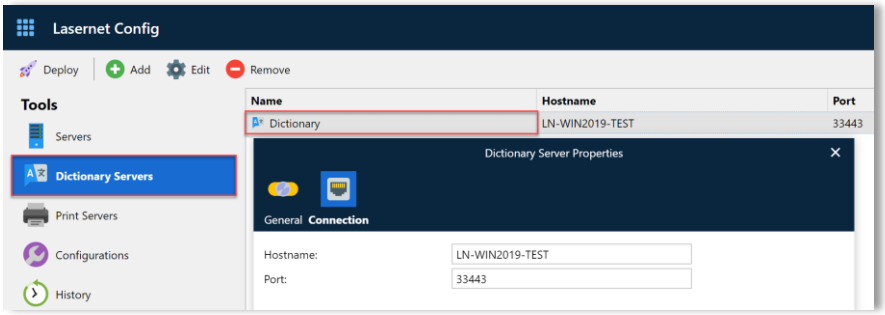
ℹ This instance does not require an additional license.



Start the relevant service in Windows services, on the Lasernet Server.

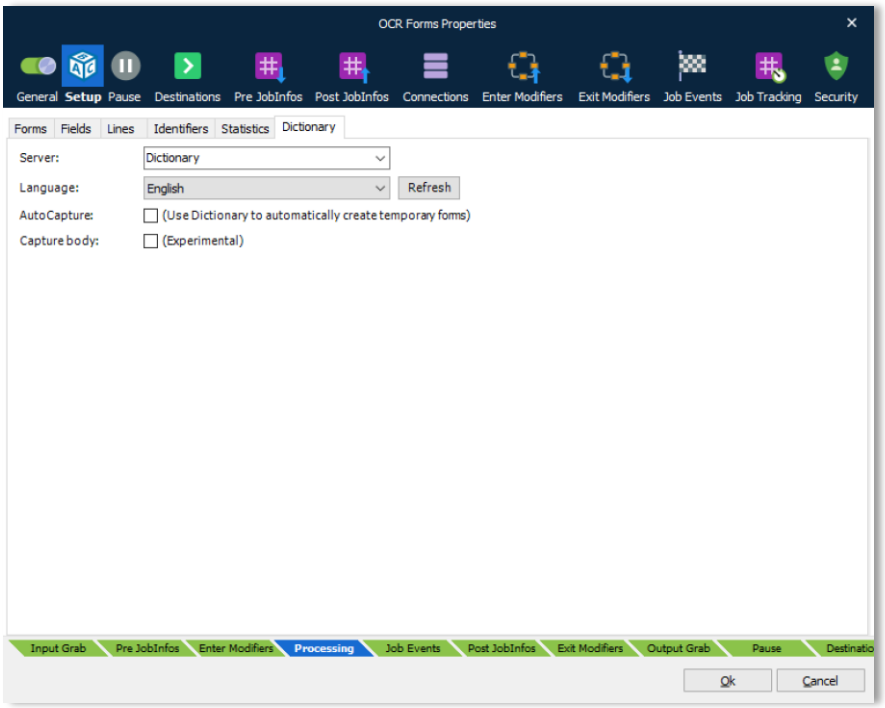


Add and configure the Dictionary Server properties for the Lasernet Config Server.



The Dictionary tab for the OCR Engine added to the Lasernet Developer has the following options:

Server From the drop-down menu, select a Dictionary Server added to the Lasernet Config Server.



Note: Leave the field empty to disable the connection to a Dictionary Server.

Language Select the primary language of the dictionary for users running the OCR Editor.

AutoCapture Select the checkbox for this feature and the OCR Engine will create temporary OCR forms for processed documents, using the dictionary server as an automated tool to extract any possible OCR field defined in the dictionary.

To improve extraction of data, the dictionary Server must be trained, by the OCR Editor, with information about OCR Fields and metadata recognition present in the processed documents.

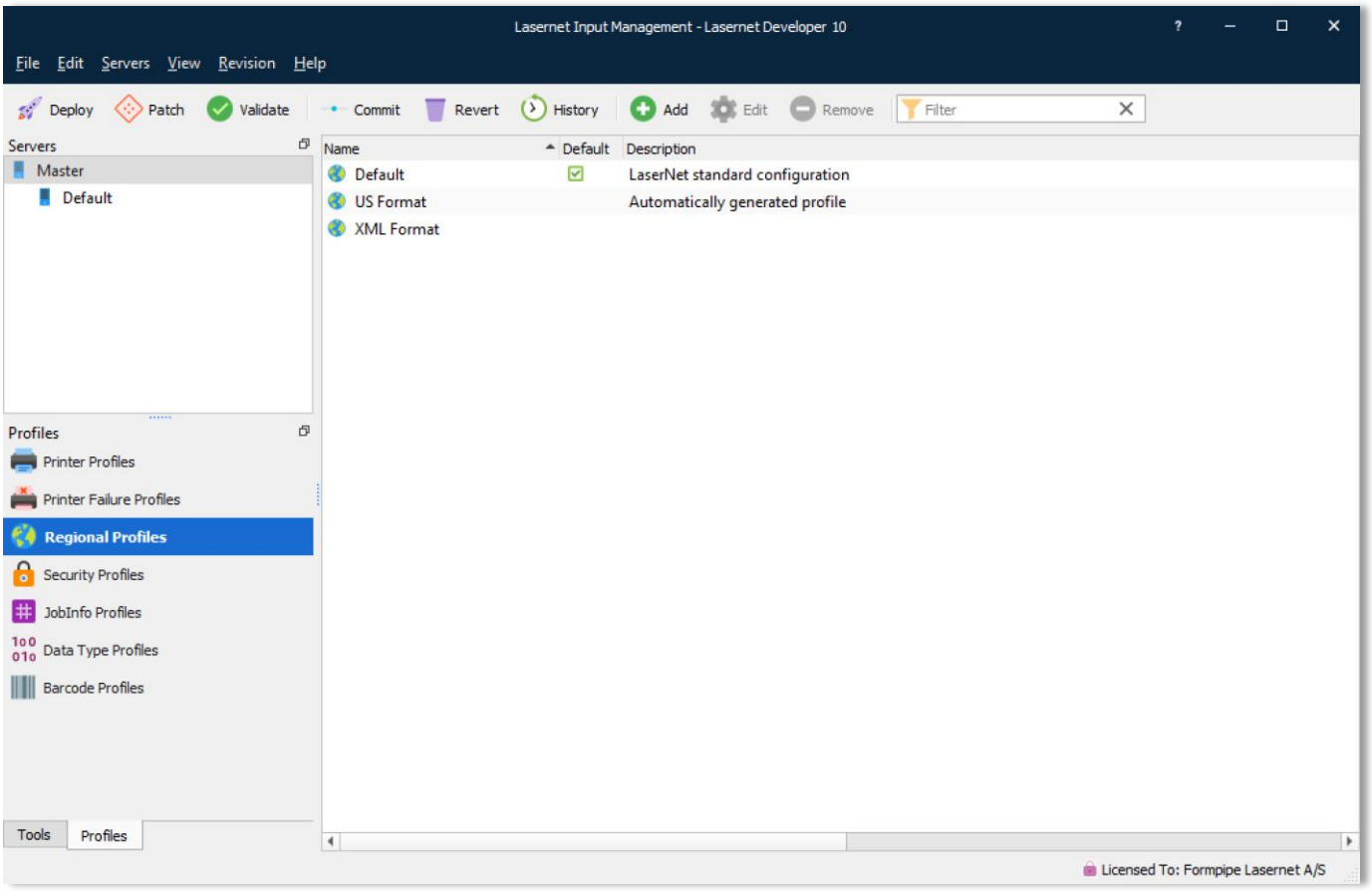
Captured documents are, by default, paused by the system and validated/released by an end user from the Lasernet Client.

Capture body: Select the checkbox for this feature to extract item lines in the body of a document.

Note: This is a complex procedure and is recommended only for documents with a simple structure in the item lines.

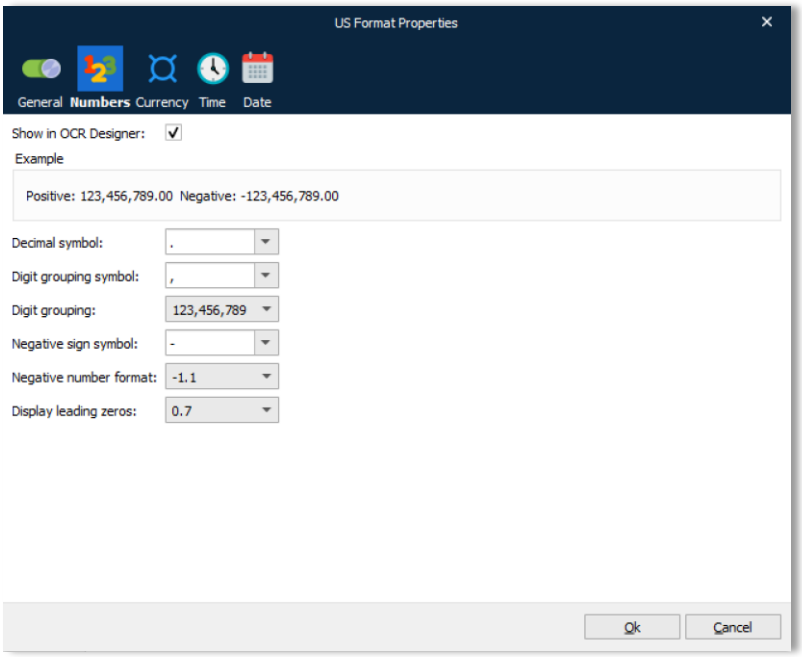
3.3.9 Field Names and Regional Profiles

When setting up unique formats for amounts and dates in the output format, regional profiles must be created based on the incoming data fields. Regional Profiles, which are defined in the Lasernet Developer, will be available in any OCR Forms designed in OCR Editor.



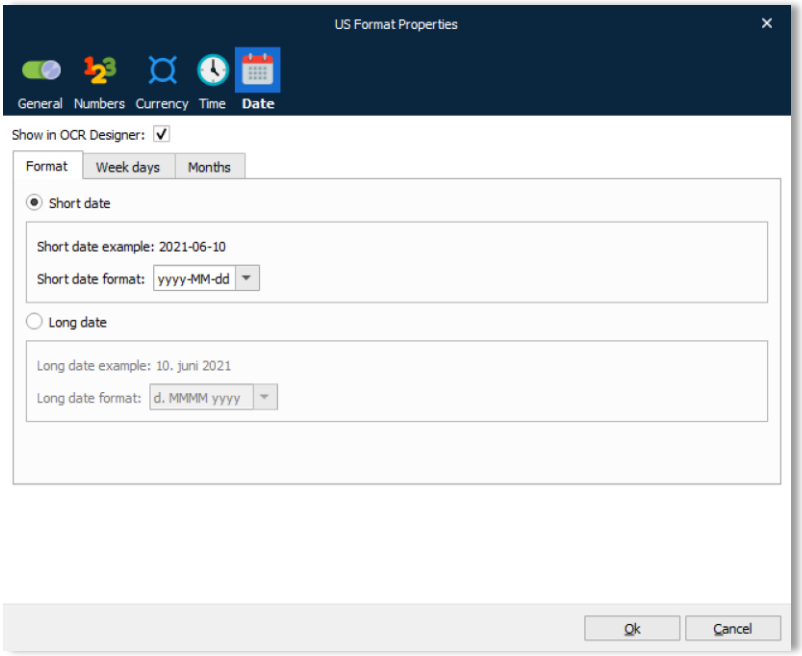
Formats for Dates and Numbers are accessed via Regional Profiles by OCR Editor.

Numbers and Date formats can be defined for any format that exists in the incoming documents and will be available for the OCR Editor.



Field types defined as Numbers in the OCR Engine will make use of Number formats defined via the Number settings in the Regional Profiles.

Select the **Show in OCR Designer** checkbox to show the Regional Profile containing this Number format.

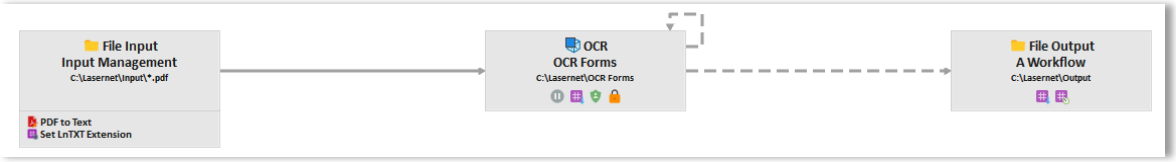


Date formats use the Date settings. Field types defined as strings do not use any regional settings.

Select the **Show in OCR Designer** checkbox to show the Regional Profile containing this Date format.

3.3.10 Example of a workflow

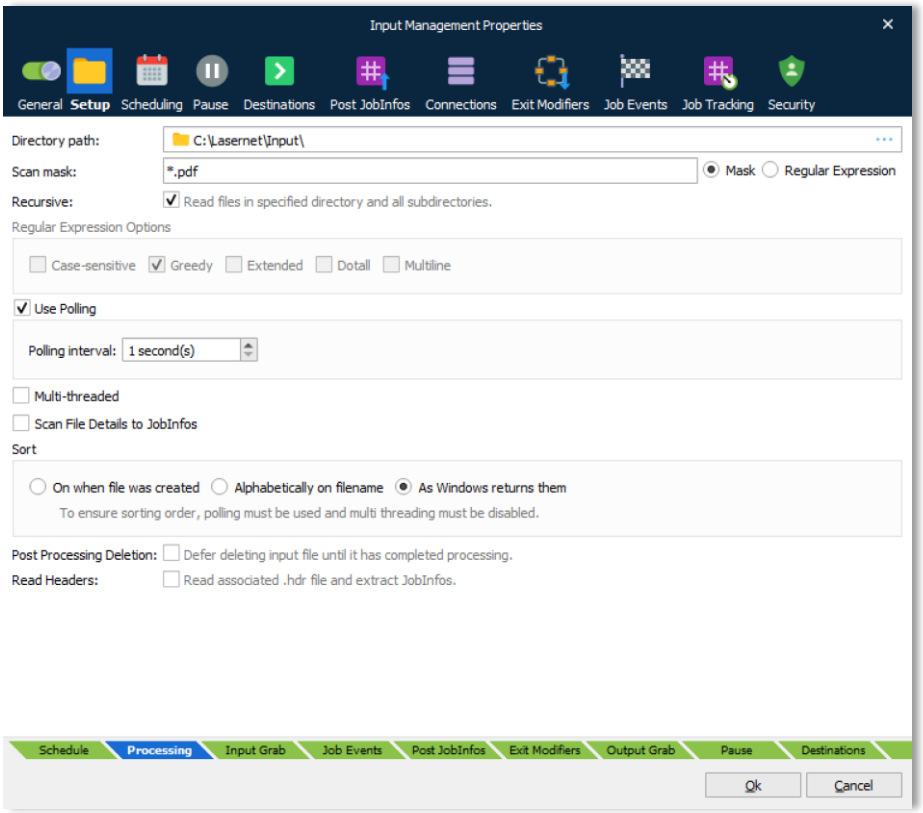
The input format for the Lasernet workflow is a PDF document. If your documents are scanned TIFF files you must configure the Tesseract OCR module to convert a scanned TIFF file into a text readable PDF document before it is processed by the PDF To Text module.



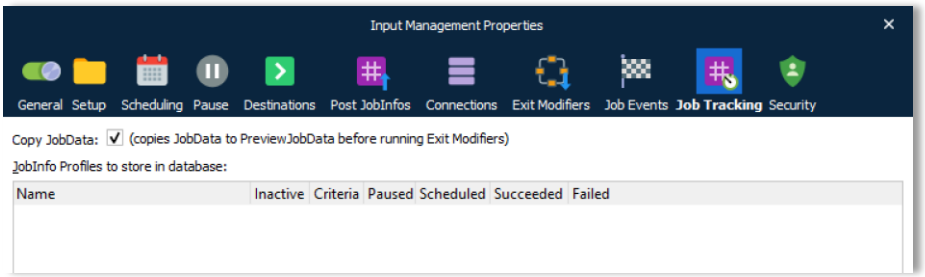
A PDF to Text modifier is used to convert PDF into a text document, which is used by the OCR Engine to extract the contents into an XML file. The output can be stored on any output location accessible to your external workflow system.

3.3.11 Input Module

Add a File Input Port and set the scan mask to read for PDF files.



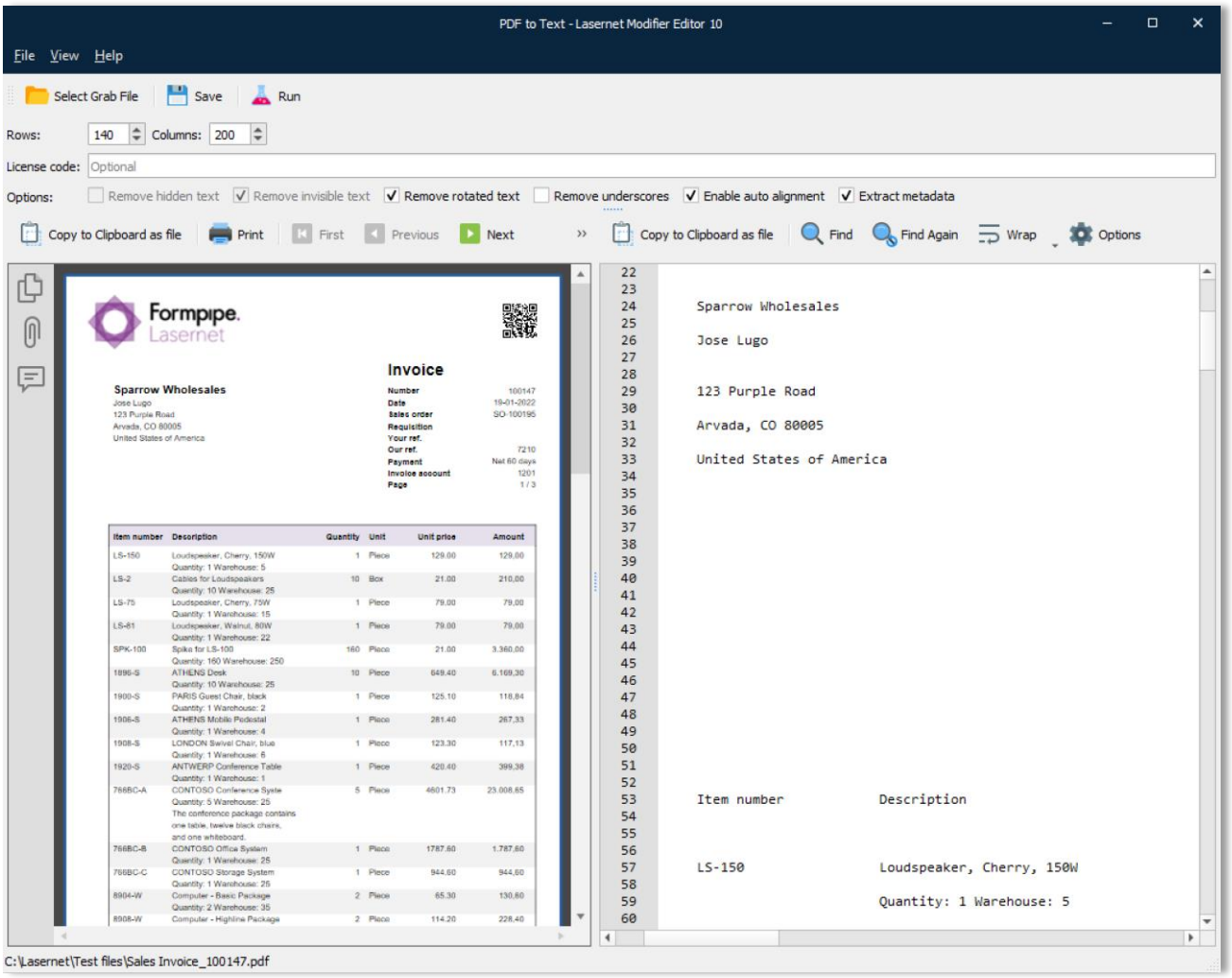
Enable Job Tracking ► Copy JobData on the input module to create a copy of the JobData before it is converted into text.



The incoming PDF document is copied into the JobInfo named **PreviewJobData**. The Lasernet Client and OCR Editor are now able to detect that the preview window shows the incoming document instead of JobData, which is a text file only.

3.3.12 Modifiers – PDF to Text and JobInfo Manipulation

Add a **PDF to Text** modifier to the File Input module to convert PDF to a text format.

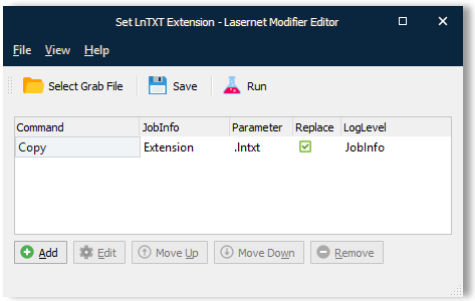


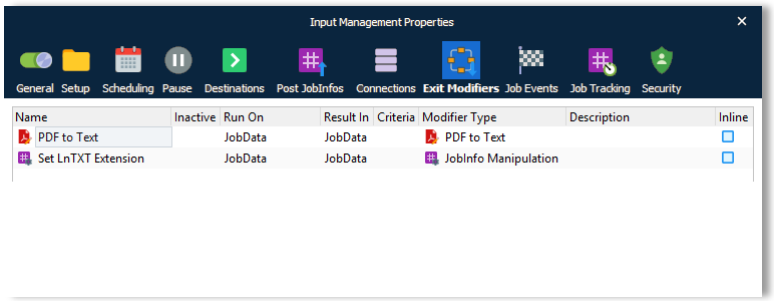
Values for **Rows** and **Columns** are used for the algorithm to extract the text strings, embedded in the PDF document, into a text grid. Rows represents the number of lines per page and columns represent the width of the text file.

In previous versions of Lasernet, an additional license code was required to run the PDF to Text modifier. To ensure backwards compatibility, you must enter your old license code here; otherwise, leave the field empty.

'Hidden' and 'invisible' text options allow you to remove text from PDF files which are usually not needed in the final text file, after conversion.

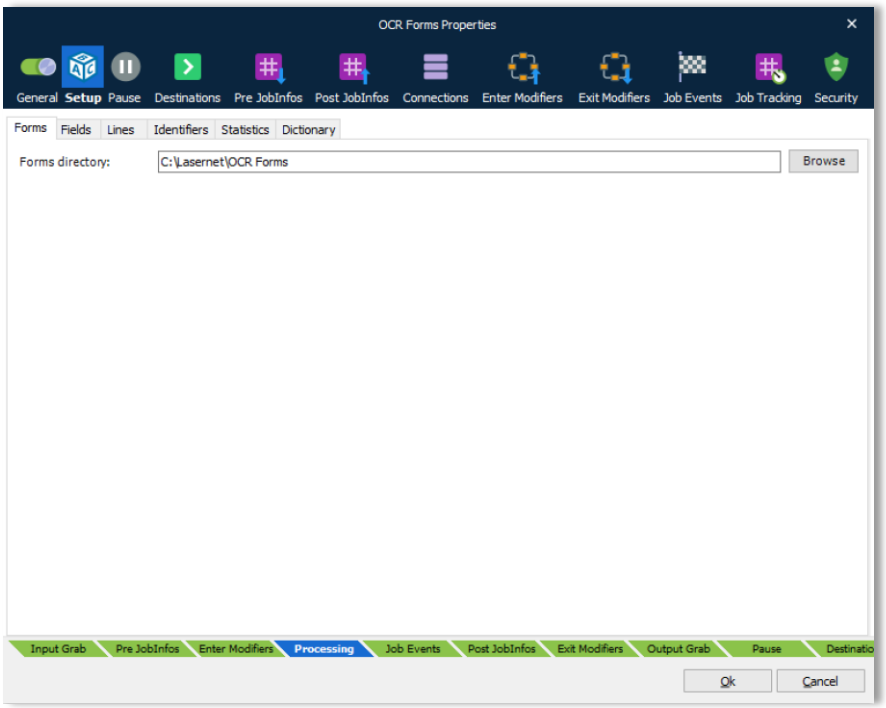
Create a JobInfo Manipulation and add the object as a second Exit Modifier on your File Input port. The modifier will associate the extension of text format with the OCR Editor.



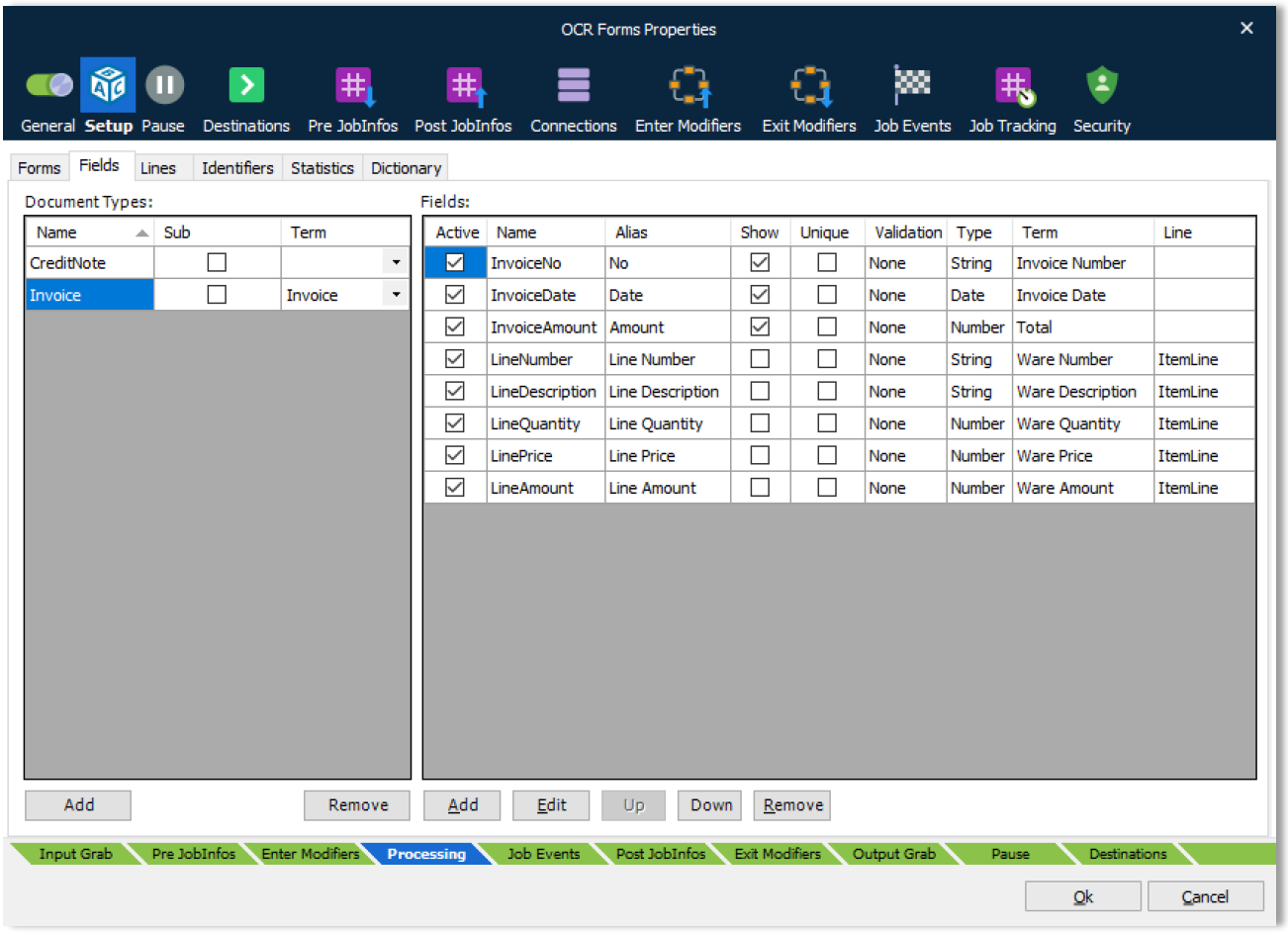


3.3.13 Engine - OCR

Add an **OCR Engine** with a forms directory in which to store uploaded OCR Forms on server site.



Set up the list of document types you want to maintain in your OCR solution and a list of Field Names and Line Types.



3.3.14 OCR Engine – additional settings

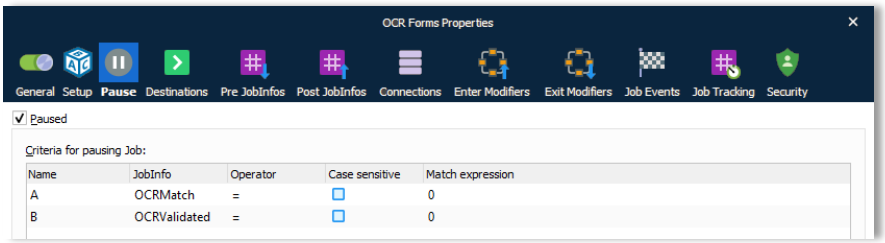
In the Pause tab you can add a pause functionality for incoming documents, which are not recognized by the OCR Engine.

For processed documents the JobInfo: **OCRMatch** will be set to either **0 = not recognized** or **1 = recognized.**

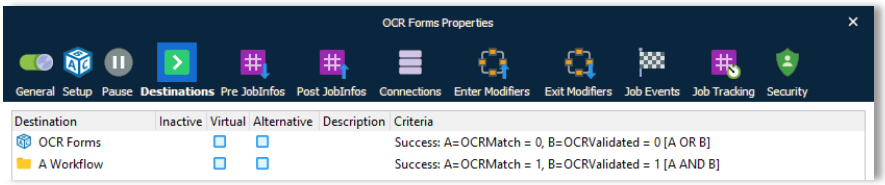
By setting up a Job Destination and a validation criterion in the Pause tab you can do one of the following:

- pause the jobs that have not been matched
- validate successful OCR Engine matches
- forward them to another module in the workflow

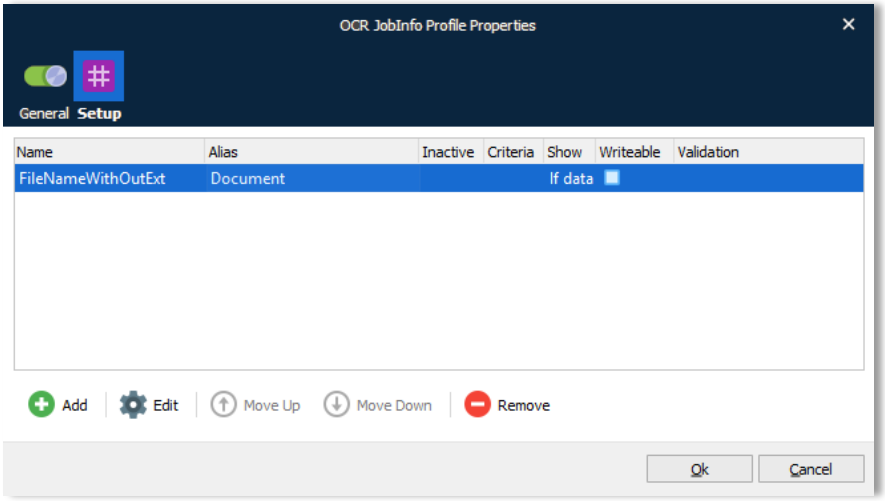
Note: OCR validation is an optional feature and if it is not configured, the JobInfo should not be added as a criterion for pausing jobs.



When the job has been processed by the OCR Engine, the JobInfos OCRMatch (Boolean) or OCRValidated (Boolean) (optional) can be used for checking if the job should be forwarded to the Form Engine or sent back to the OCR Engine.

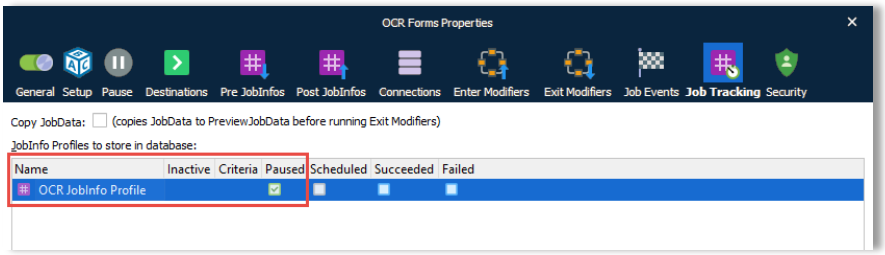


When saving jobs in pause mode, you can save other available JobInfos as metadata for the job. This is maintained via a JobInfo Profile.



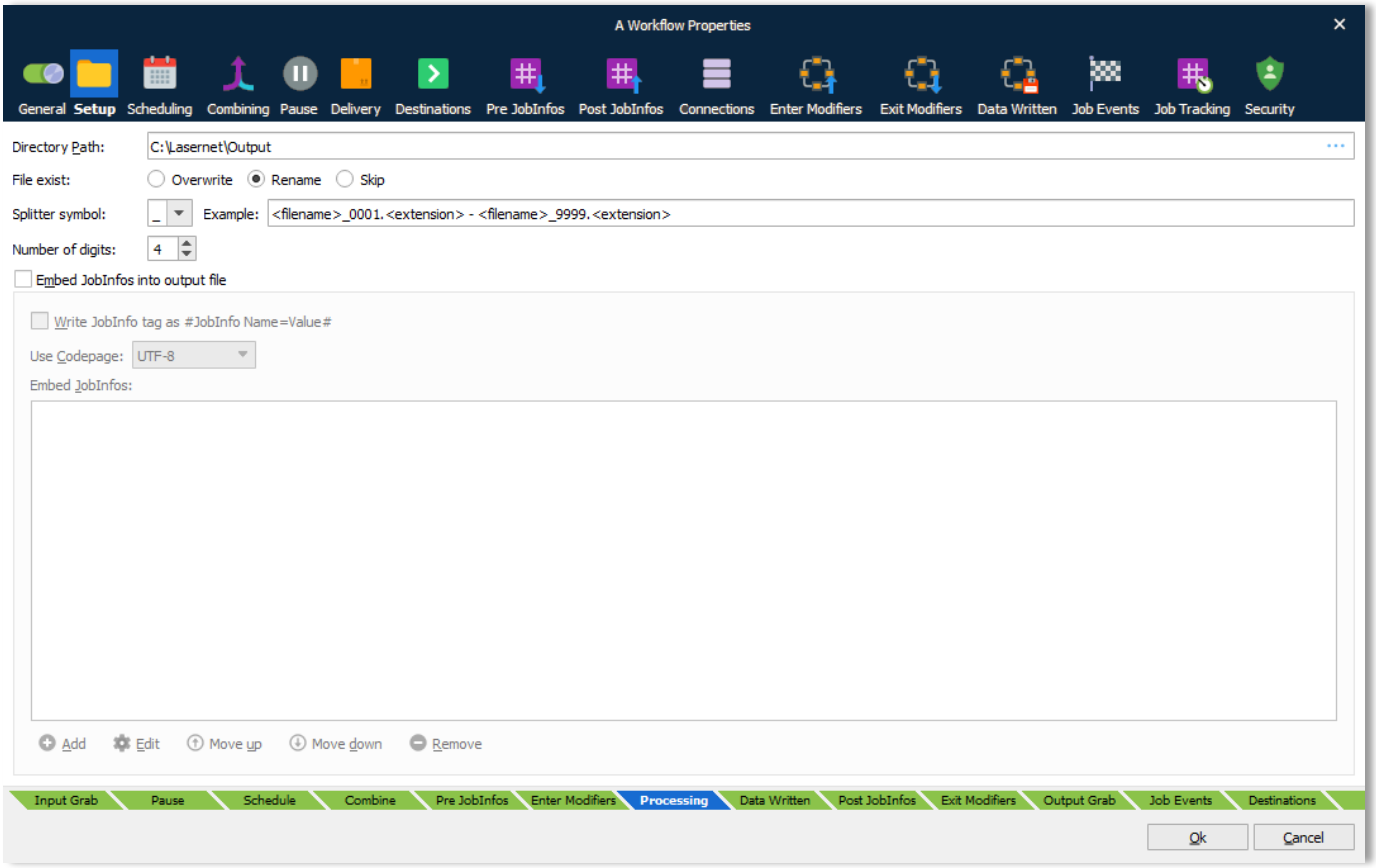
In this example, a JobInfo Profile named “OCR JobInfo Profile” is added for storing additional metadata like the document name without the file extension (FilenameWithoutExt) to a database.

When the JobInfo Profile is created you can assign the profile to the OCR Engine ► Job Tracking tab and activate it for paused jobs only.



3.3.15 Modules – Output

Add a File Output module to save the data extracted by the OCR Engine. This must be in a location accessible from your external workflow system.

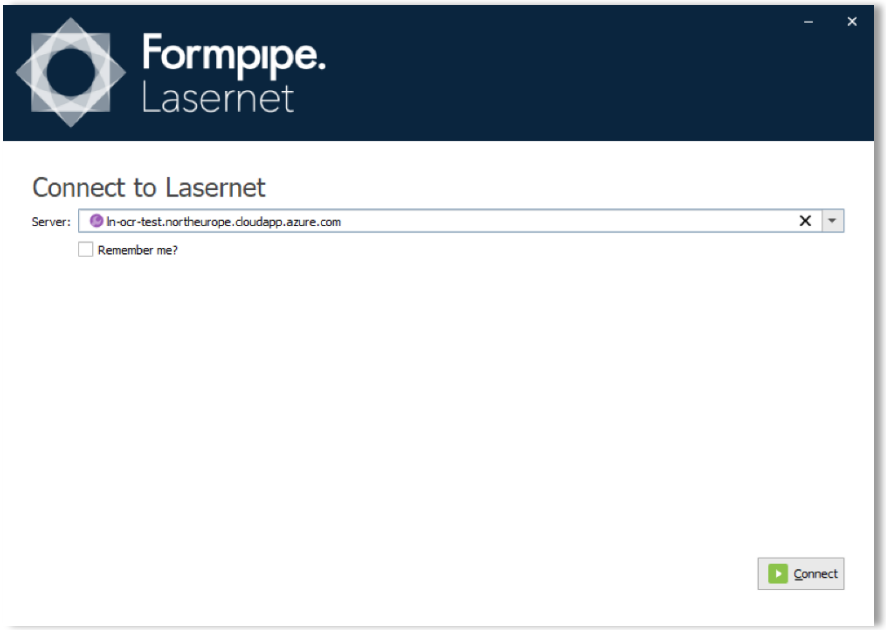


Set whether you want to overwrite, rename or skip the output if a file already exists with the same file name.

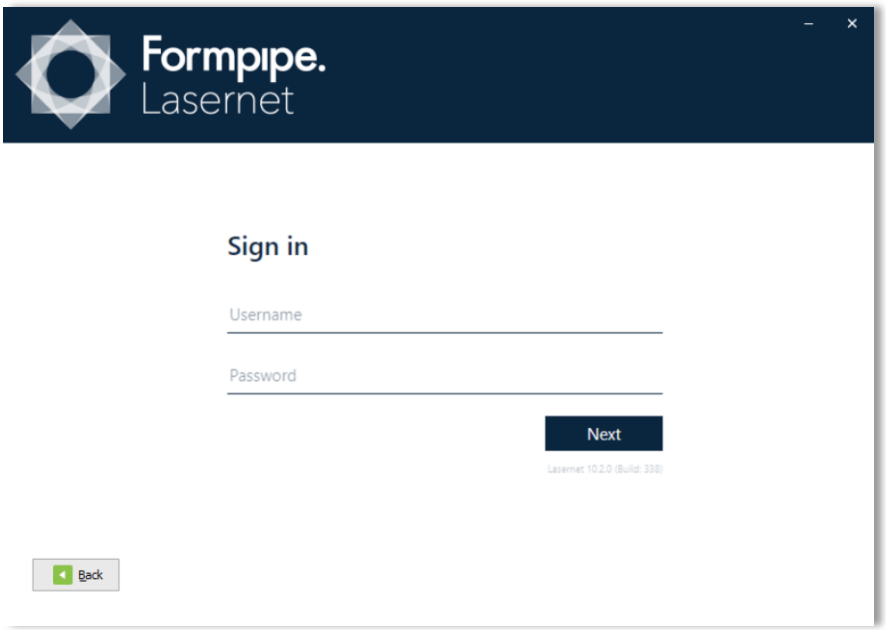
3.4 Getting started – Lasernet Client

To open a paused job saved by the OCR Engine, the Lasernet Client needs to be installed on the same computer as the OCR Editor.

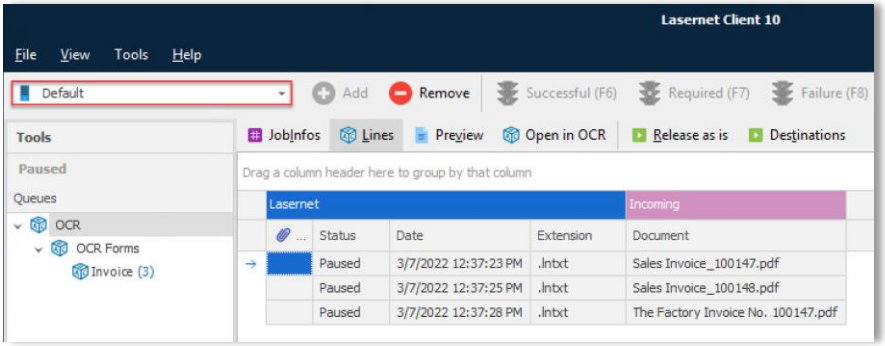
Before you can retrieve the paused job, connect to a Lasernet Config Server.



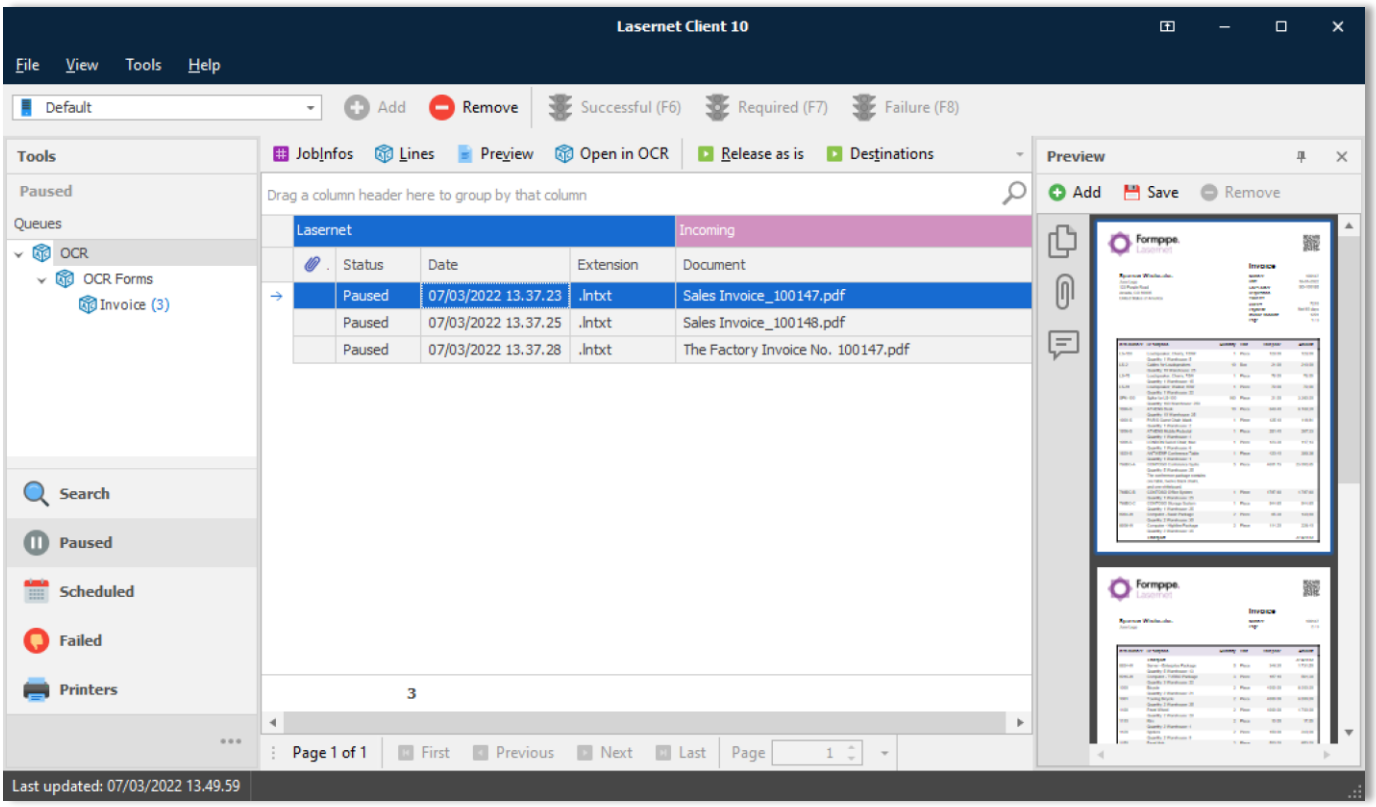
Sign in with your credentials.



Select a server instance from the Lasernet server.



You can now view the paused jobs stored by the OCR Engine and edit them in OCR Editor.

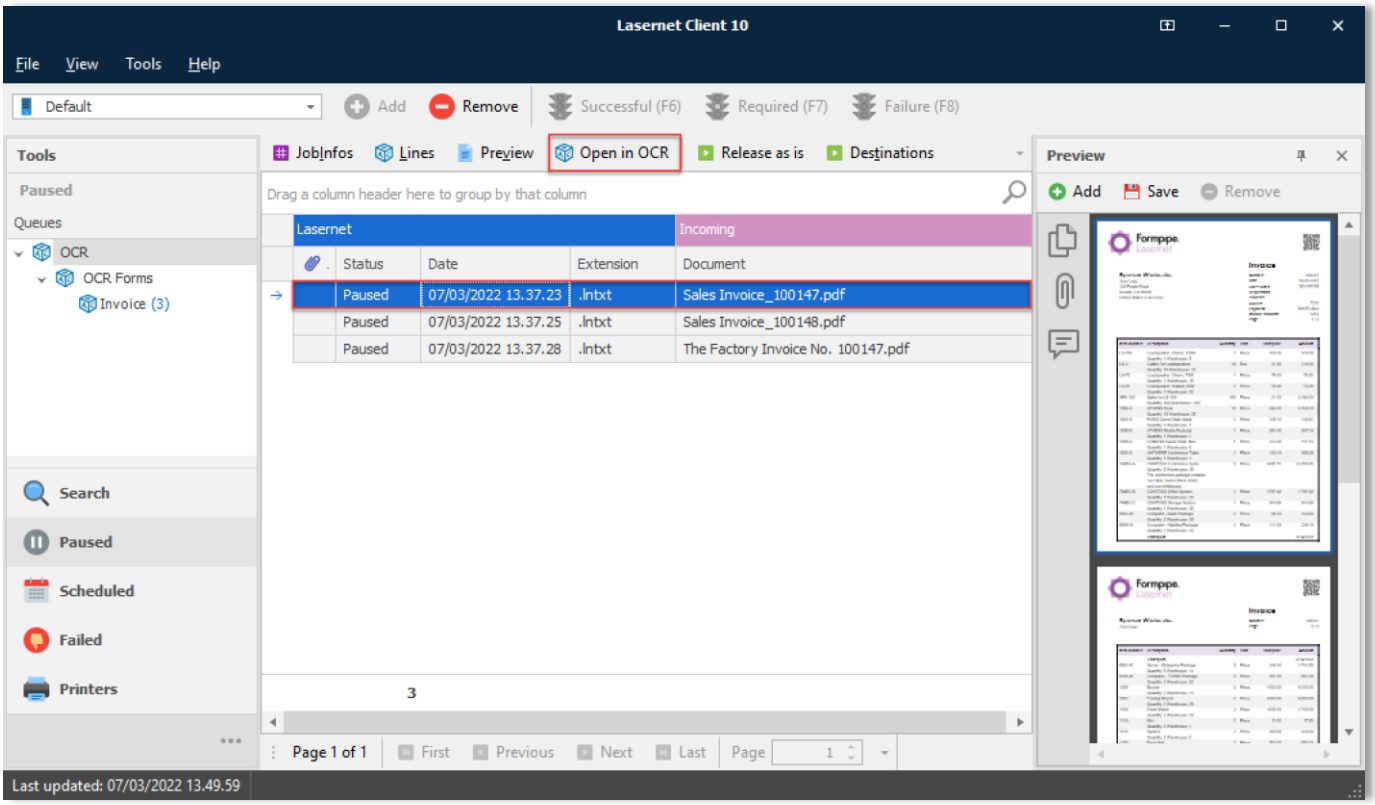


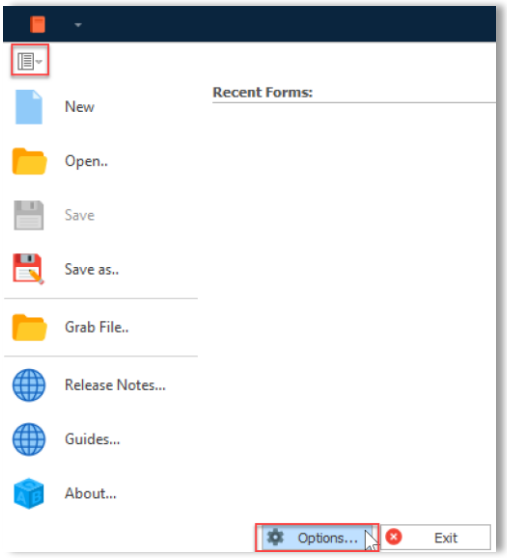
Select Paused documents and click **Open in OCR** to create an OCR form.

If you want to retry/release a paused job, first select the row, then right-click on it and choose **Release as is**. If the job is recognized by the OCR Engine (for example, if you have finished setting up an OCR form for it), it will be removed from the queue and forwarded to a Job Destination added to the OCR Engine in your Lasernet configuration.

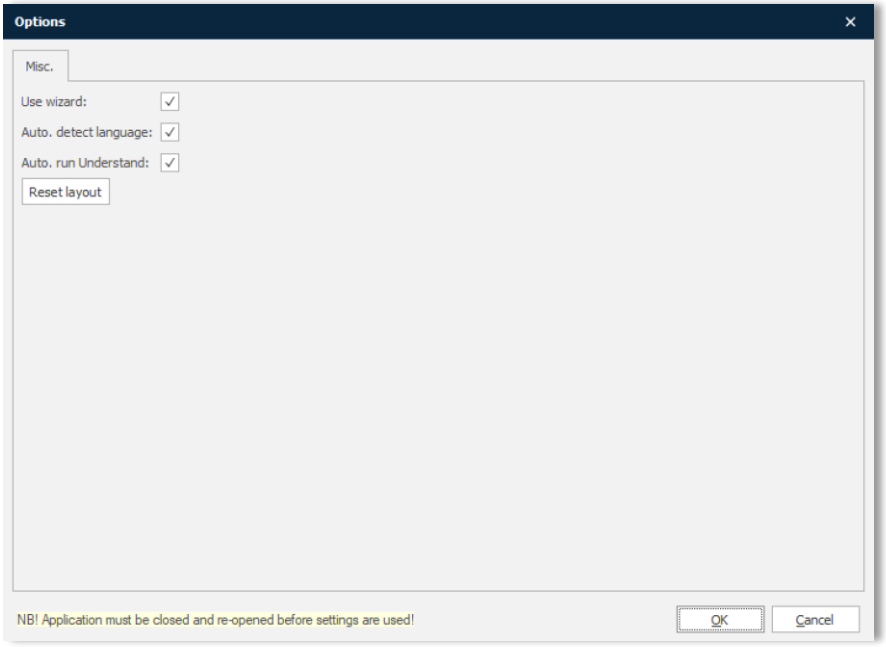
3.5 Getting started - Lasernet OCR Editor

Start the OCR Editor by selecting a document that is already paused in the Lasernet Client (if any) and click **Open in OCR**.





When you have logged in to the OCR Editor, go to **Options ► Misc.** tab and activate your preferred tools.



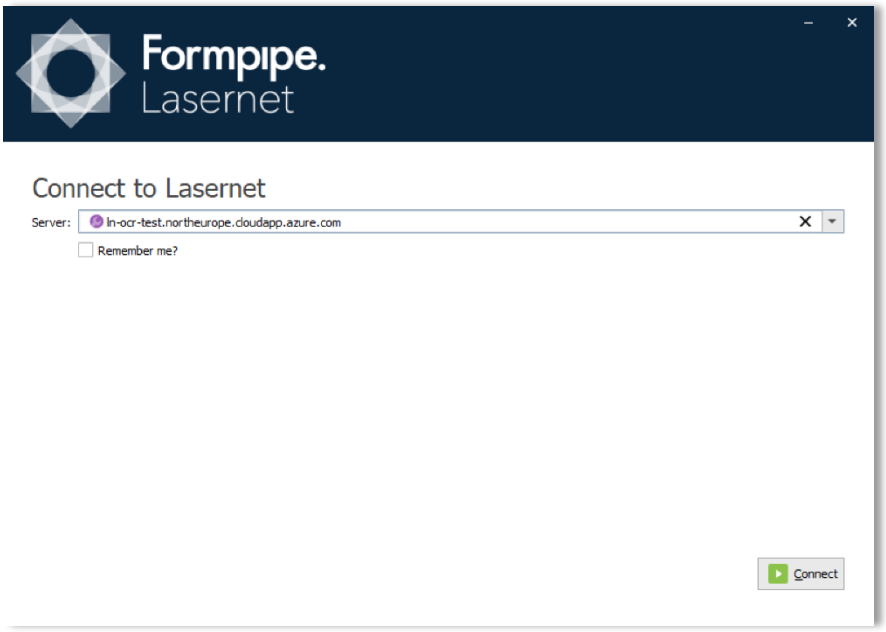
Wizard is a tool that will guide the user to capture OCR Fields in documents via a built in wizard.

Auto. detect language will analyze the language of the document and automatically set a regional profile for date and number formats, if found.

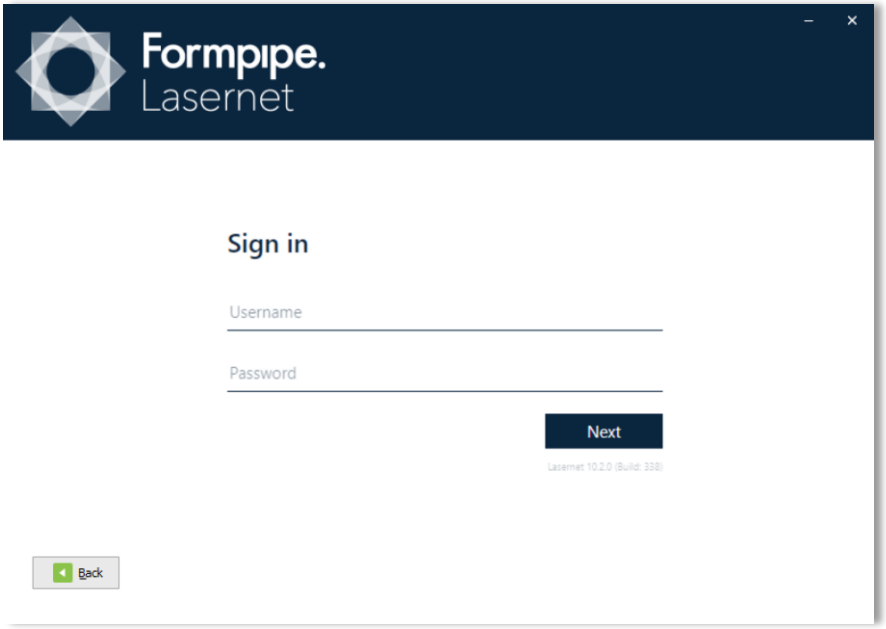
Auto. run Understand will auto capture all possible OCR Fields, as the first step when a document has been opened by the OCR Editor and if the dictionary is activated on the Lasernet Server.

3.5.1 Open OCR Editor direct

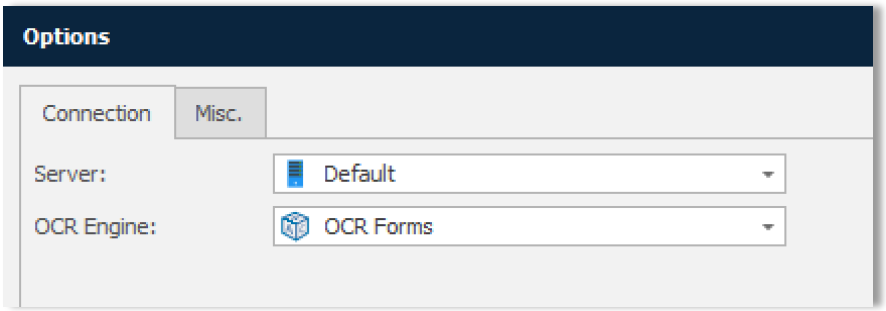
Starting the OCR Editor, without the Lasernet Client, you must first Connect to Lasernet Config Server.



Sign in with your credentials.



Select a server instance and an OCR Engine.

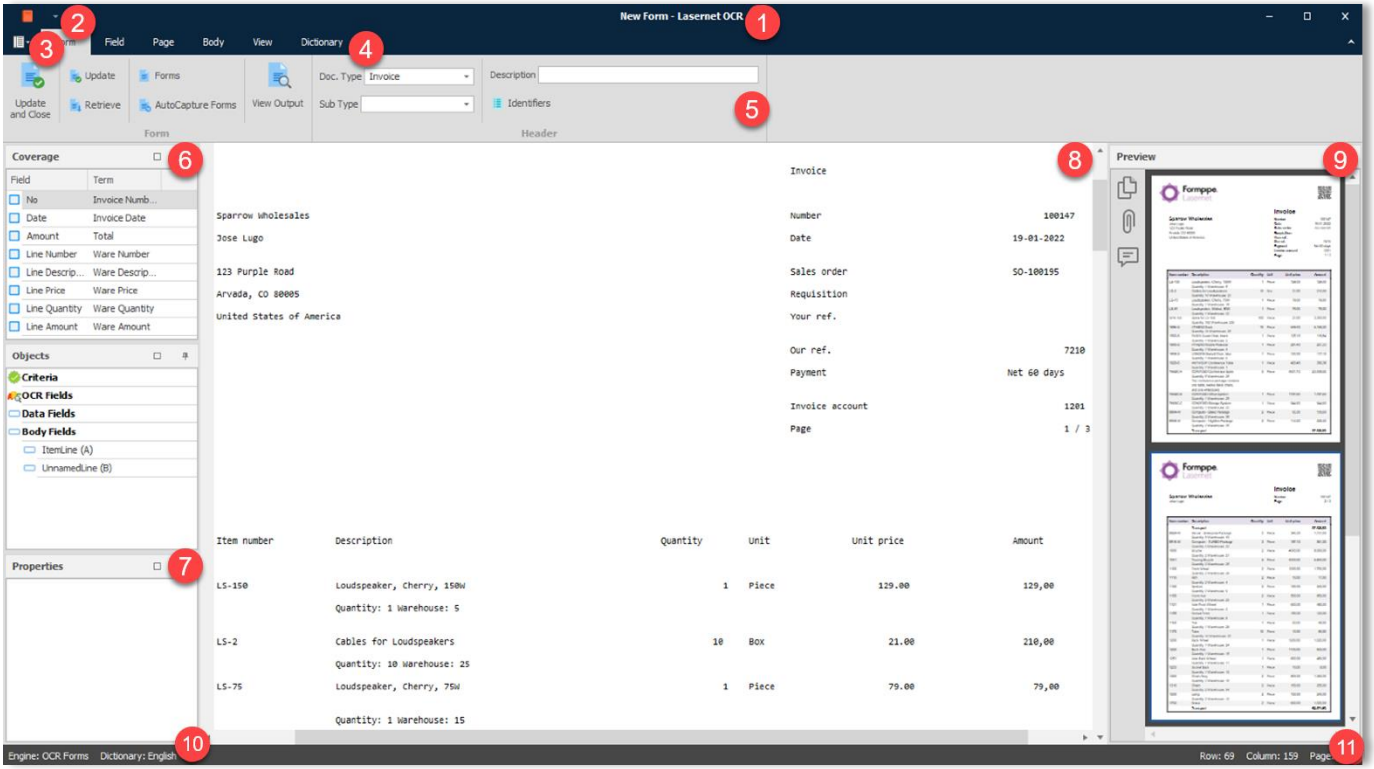


Server The server name on which the Lasernet Server is running.

OCR Engine Logical module name defined for the OCR Engine. Multiple OCR Engines are supported on the same server.

3.5.2 The user interface

The user interface consists of a window showing text data extracted from PDF or TIFF documents. Only the input view is needed for OCR Editor because the output format is already pre-defined (compared to the Lasernet Form Editor, where both the input and output format must be defined manually).



Form name (1)

The name of the form. If no name has been given, it will be named “New Form”.

Dictionary (2)

Enter Dictionary mode. Add and manage terms and aliases to understand forms.

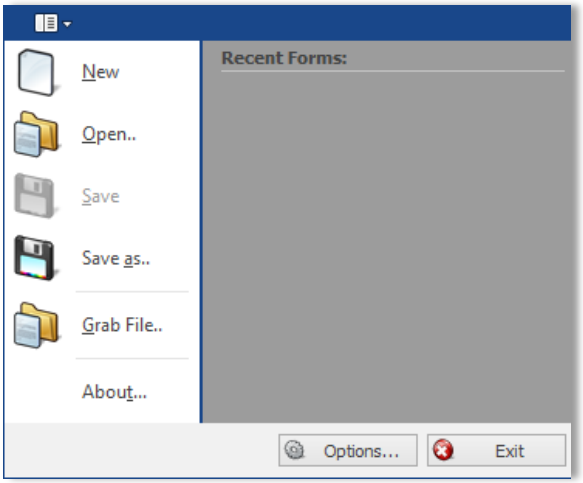
Options (3)

Set the Options for the OCR Editor – edit connection and user credentials, enable/disable the wizard for creating OCR Fields.

Menu bar (4)

You can access all the application functions and options through the various menu headings.

From the File menu you can Open, Save and create New OCR Forms. Grab files can be loaded into an existing form and the Options for the application can also be set.

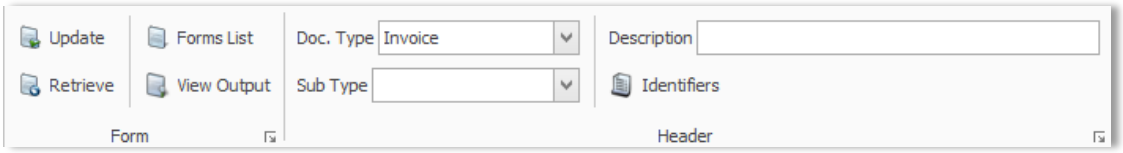


Recent forms can be opened from the Recent Forms list.

If you want to load a form that has been saved on the server, we recommend that you open it via the Form ► Retrieve functionality. This is because the path is not shown in the Recent Forms list so it is possible to accidentally use an old, temporary or locally-stored version of the file instead.

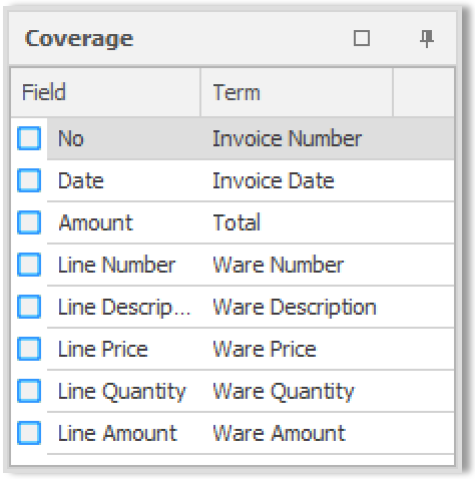
Menu settings (5)

Settings for form, field, body and view options.



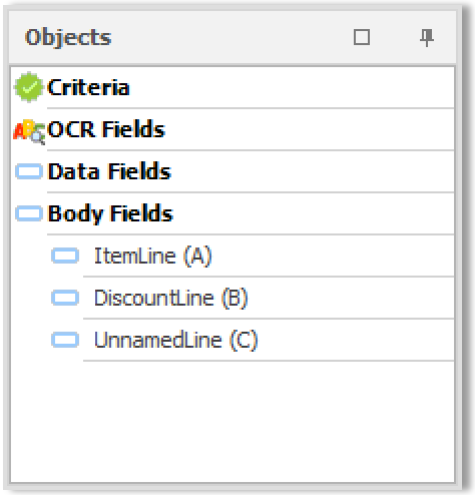
Coverage (6)

A list containing available Data Fields for the document type defined in the OCR Form. Activated Data Fields indicate that the field is already included in the OCR Form.



Object list (7)

A list containing objects added to the form. The list includes Criteria, OCR Fields, Data Fields and Body Fields.



Design View (8)

Document (JobData) represented in text format.

Preview (9)

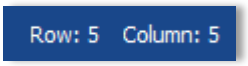
A preview of the incoming PDF document (contents of the JobInfo PreviewJobData if it exists).

OCR Engine and Dictionary (10)

The name of the OCR Engine to which the application is connected, and the default dictionary language.

Row and column (11)

Positioning for selected object.



3.6 Getting started – OCR Forms

3.6.1 Creating a Form

OCR Forms cannot be designed without having business data to work with. This data needs to be stored in a text file in a shared location that is accessible to the OCR Editor, or can be opened from a paused queue in Lasernet Client.

You can create a new OCR form by following the relevant steps:

- **Scenario 1** – Lasernet has converted and saved a text copy in a shared location.

  Start OCR Editor. Go to **File ► Grab file** and browse for a valid text file.

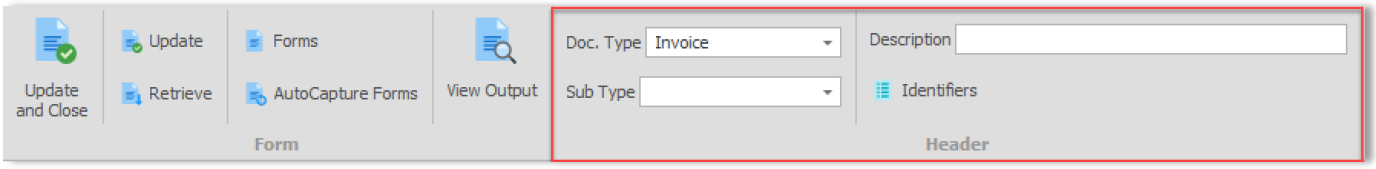
- **Scenario 2** – Lasernet has converted and saved a text copy in the internal database.

  Start Lasernet Client. Select the Paused tab and double-click on any of the documents listed for the OCR Engine. The OCR Editor will open the document.

ℹ When the text / grab file has been loaded you are ready to start creating an OCR Form.

3.6.2 Form Properties

In the Form properties you are able to manage the properties of your form.

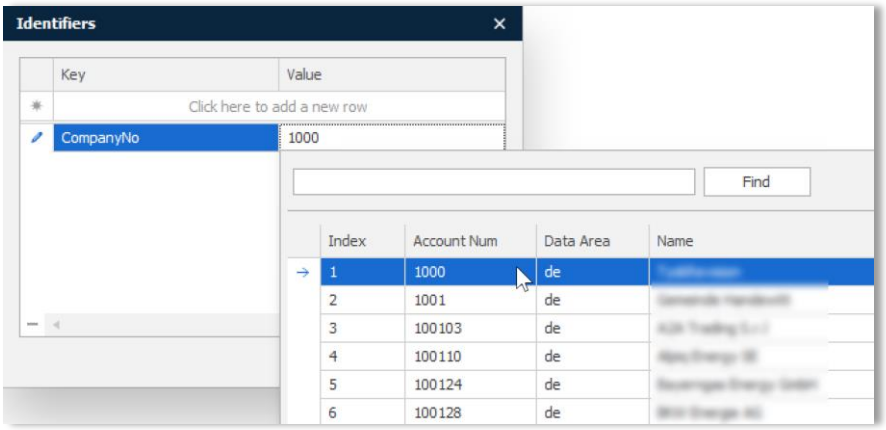


Doc. Type Type of document. List retrieved from the OCR Engine. Field names belonging to this document type will be available in the form.

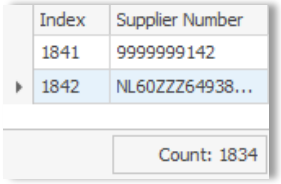
Sub Type Sub type for document. List retrieved from the OCR Engine. Field names belonging to this sub type will be available in the form.

Description: Description of the form. The text will be added as a value in the final output.

Identifiers: Identifiers are fixed field names and values that can be added to output data. Only characters from A–Z are recommended as valid key names.



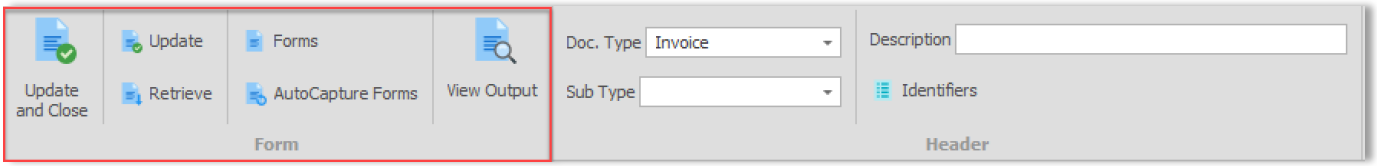
⚠ If the CSV file contains a parsing error due to an incorrect CSV format, a dialog will be displayed with a warning message and the line number saying: “Error parsing CSV file. Records with incorrect formatting will not appear in the list, but are still allocated an Index number.



If there is mismatch between the highest index number and the total count number, please report the error to an administrator of the Lasernet Config Server. This can be caused by a corrupted CSV file.

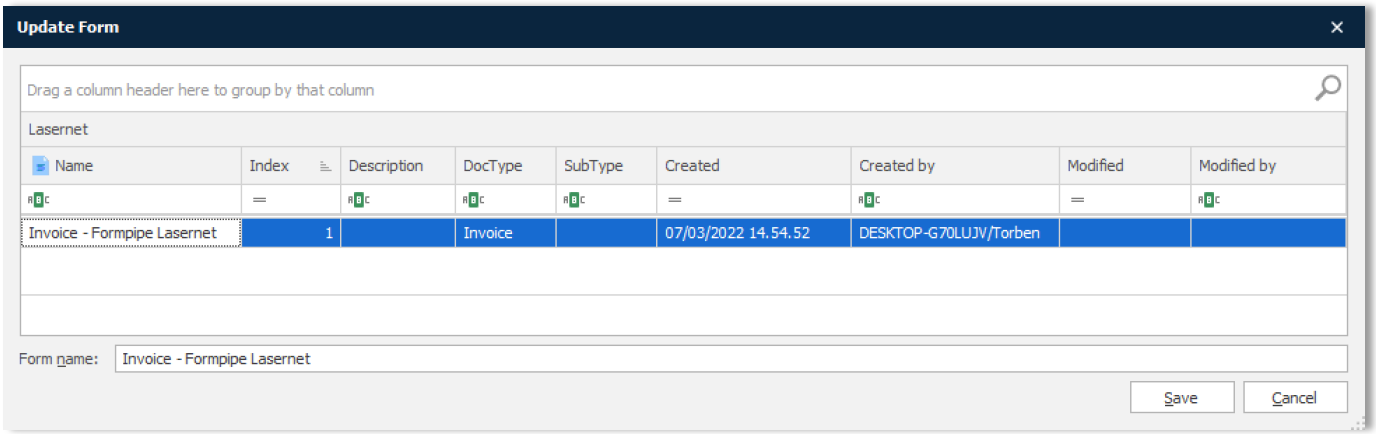
3.6.3 Form Tools

Administrative tools are available in the ribbon bar. You can update, retrieve, administrate and view OCR Forms.



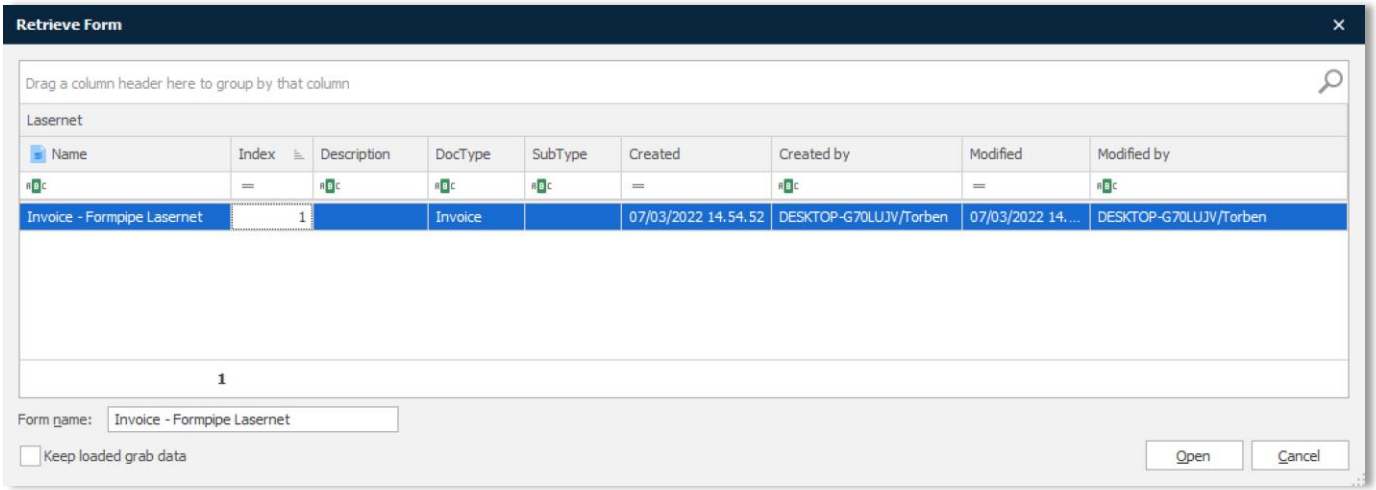
Update and Close See information about Update. The application will close after clicking.

Update Saves the OCR Form in the OCR Engine. Before an update, you must ensure that the OCR Form contains unique criteria. The OCR Engine will parse through the OCR Forms in the list index order. The first form with a criteria match will be processed. The order of the index is managed from the Form Lists tool.



You can either save the OCR Form as a new document or overwrite an existing one.

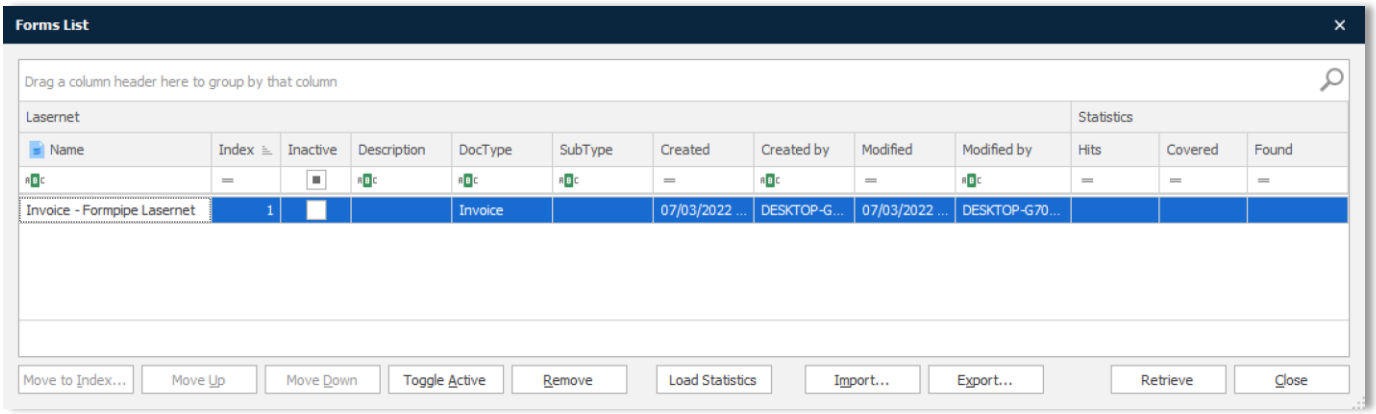
Retrieve Retrieves an existing live OCR Form from the OCR Engine.



Click an OCR Form in the form list and you will be able to view and edit the settings. A form that is already open must be saved before retrieving a new form.

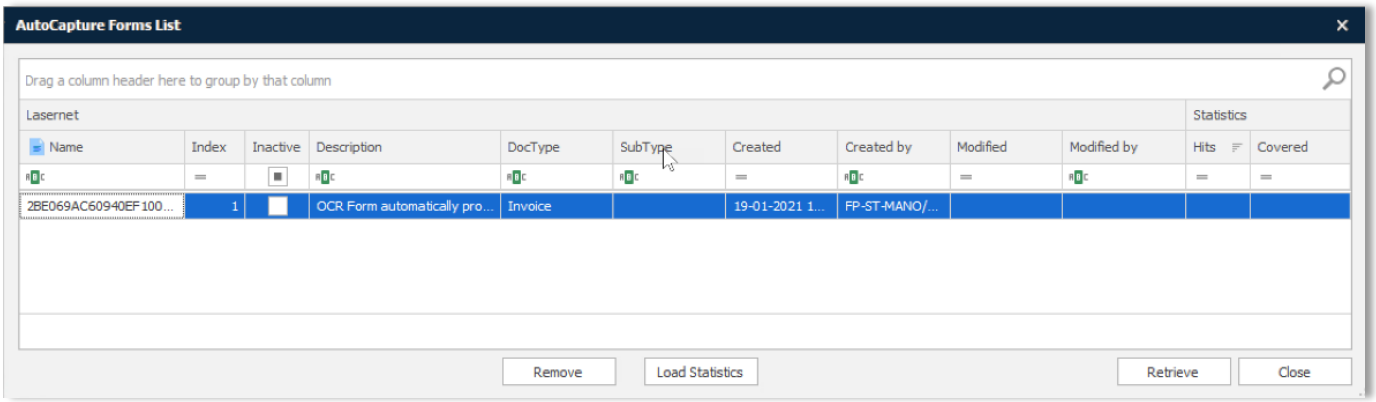
Select the **Keep loaded grab data** checkbox to use loaded grab data instead of last used grab data stored within the form at the last update.

Forms Lists the active OCR Forms already updated on the OCR Engine and running in live mode.



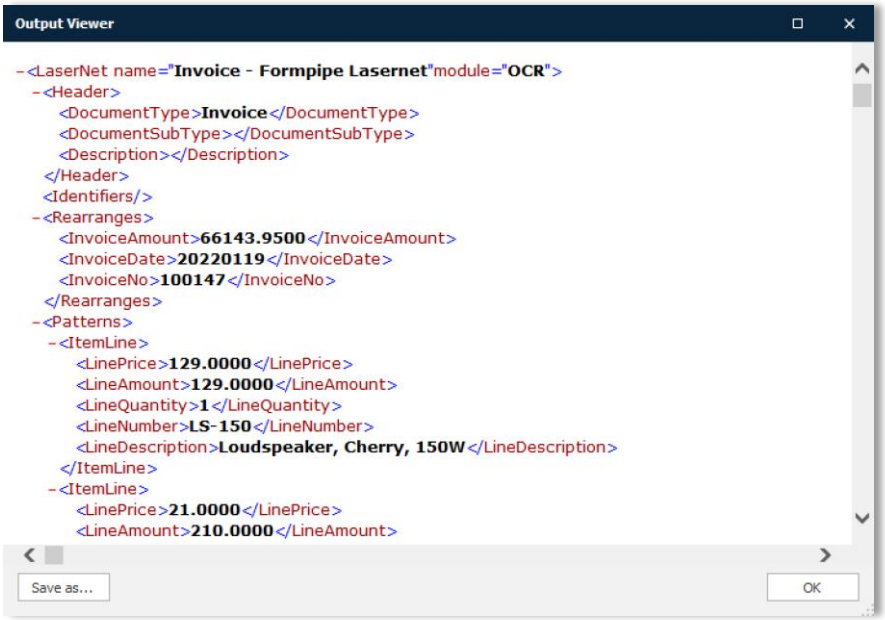
Forms can be moved up and down to change the recognition order or deleted from the form list in the OCR Engine. Forms are deleted immediately and removed from the live server. To save a copy, you must retrieve the form and use the **File ► Save as** function.

AutoCapture Forms A list of OCR forms automatically captured by the OCR Engine.



Opening auto captured forms makes it possible to **Remove** them or **Retrieve** these and edit and save them as real OCR forms.

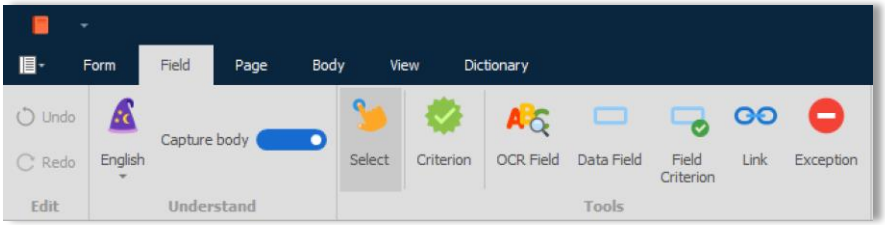
View Output View a preview of the final output, in XML or Text mode according to the setting on the OCR Engine.



Click **Save as…** if you want to save a local copy of the XML. The output file can be opened in the Lasernet Form Editor or any other XML editor.

3.6.4 Field Properties

The OCR Editor toolbar includes various tools for defining Criteria, OCR Fields, Data Fields, Link Fields and Exceptions into the OCR Form output format.



Understand To enable the Understand button, a dictionary must be available. Click the arrow icon and select a language from the drop-down menu. Click **Understand** to automatically create OCR Criteria, OCR Fields and Data Fields for words / labels in the form that match words found in the dictionary. By default, the solution is delivered with an English, Danish, Swedish, Norwegian and Finnish dictionary. Other languages must be maintained by the end user.

Capture body Slider to turn on and off the capture of item lines, in the body of the document, when clicking the Understand button.

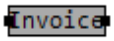
Select Select an existing object by clicking it.

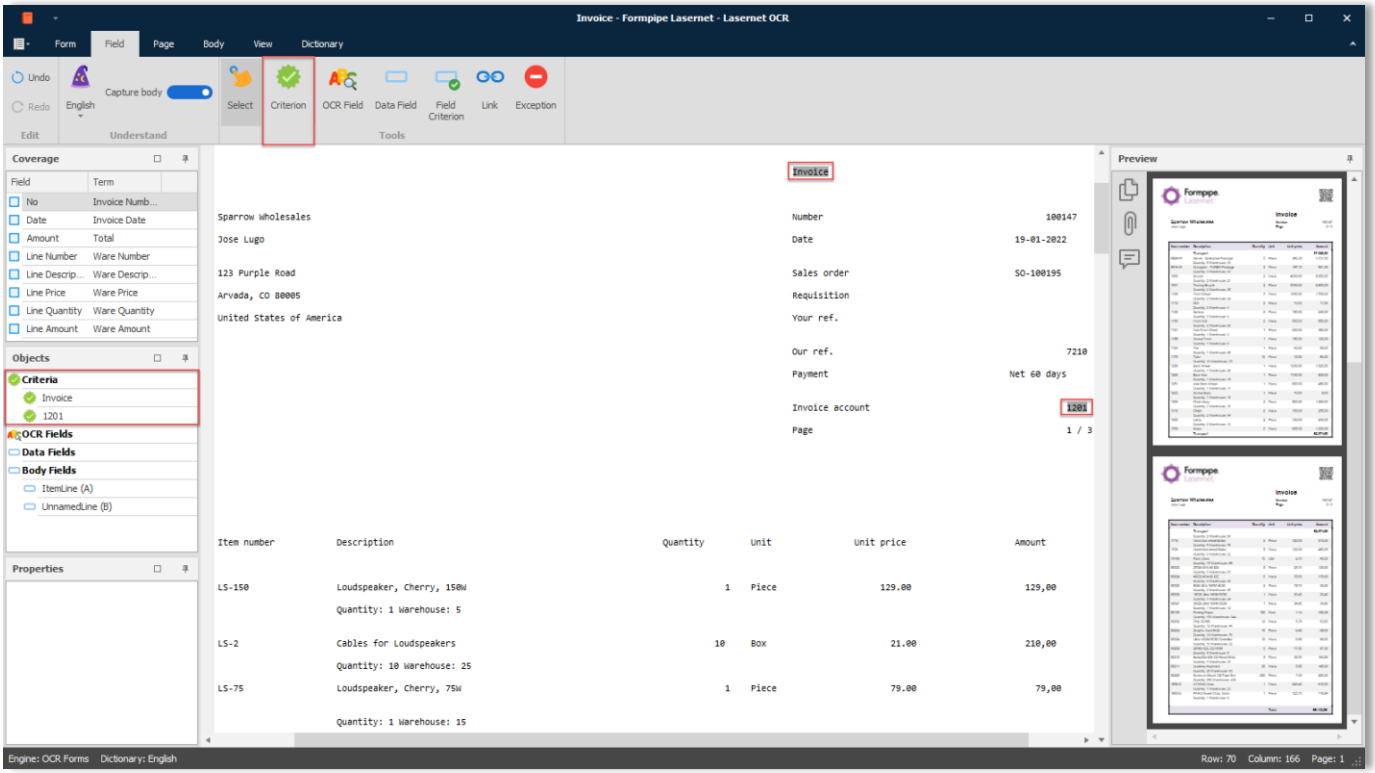
Criterion	Manually select a criterion to tell the OCR Engine to scan for specific data strings and match the form. The full page will be scanned to detect the text string in any position. We recommended that you create several criteria for each OCR Form to ensure more accurate recognition. For example, document type, company name etc.
OCR Field	Detects unique text strings in the document. An object will be created for the first text string in the document matching the OCR Field string (used in combination with Data Field and Link tool).
Data Field	When marking a text string in the OCR Form, the positioning for the Data Field is fixed (by default) to a defined row and column. See section 3.7.4 for more information. Several Data Fields can be linked to one OCR Field.
Link	By using the Link tool you can draw a dotted line between an OCR Field and a Data Field, to define the relative positioning of the two objects. The Data Field will then be grabbed according to its relative position from the OCR Field.
Exception	Used in the body area as an exception for excluding patterns with specific text symbology. Exceptions use ‘CONTAINS’ as an operator when matching strings. Exceptions in the object list can be edited by pressing **F2**.

3.6.5 Selections

Each tool has a selection value showing the settings of the object.

Criterion A Criterion is marked as a dark grey object:

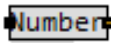


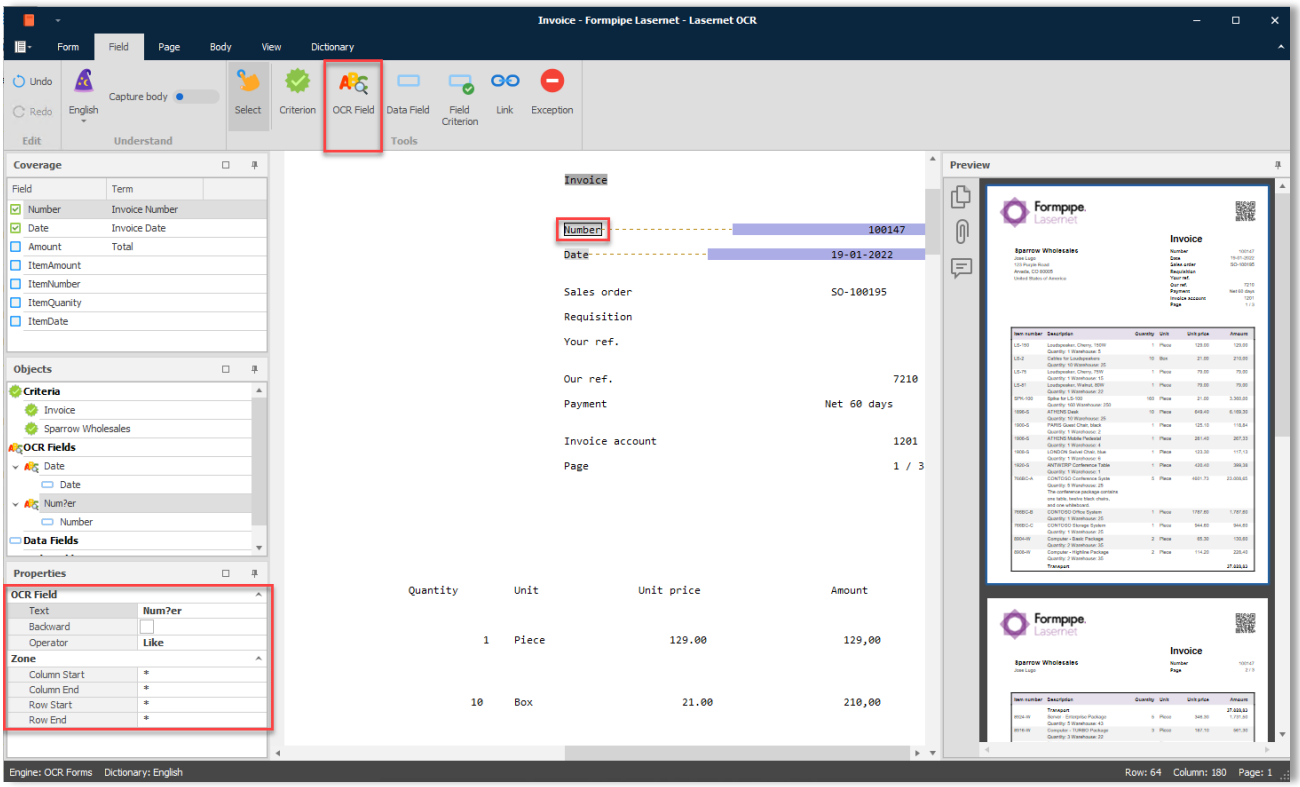


When selecting one of the objects, the selected field will show the value of the criterion in the following field:

Text Value of a criterion.

OCR Field An OCR Field is marked as a light grey object:





When selecting the object, the selection field will show the value of the OCR Field in the following field:

Text

Value of an OCR Field. The value is used to find a text on a page and link the value with a Data Field. The value can contain an exact match and wildcards for pattern matching.

The pattern matching features allow you to match each character in a string against a specific character, a wildcard character, a character list, or a character range.

Common wildcard characters in a pattern match are:

?	Any single character
*	Zero or more characters
#	Any single digit (0–9)
[charlist]	Any single character in charlist
[!charlist]	Any single character not in charlist
[A-Z]	Any uppercase character between A-Z
[a-z]	Any lowercase character between a-z

Note: Running a pattern match and the Like operator will have an impact on performance. We recommend using an Exact operator match if possible.

Backward

If selected, the search direction is from bottom upwards. We recommend this for fields that are always present at the bottom of a document, for example, the total amount.

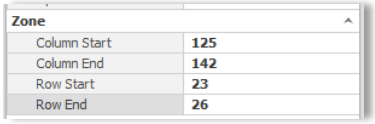
Operator

Default value is set to “Exact” where wildcard matching is not supported. Alternative value is “Like”, which works in combining with wildcards in the text value.

Zone

Per default a wildcard * defines that a search for an OCR Field runs for all pages in the document.

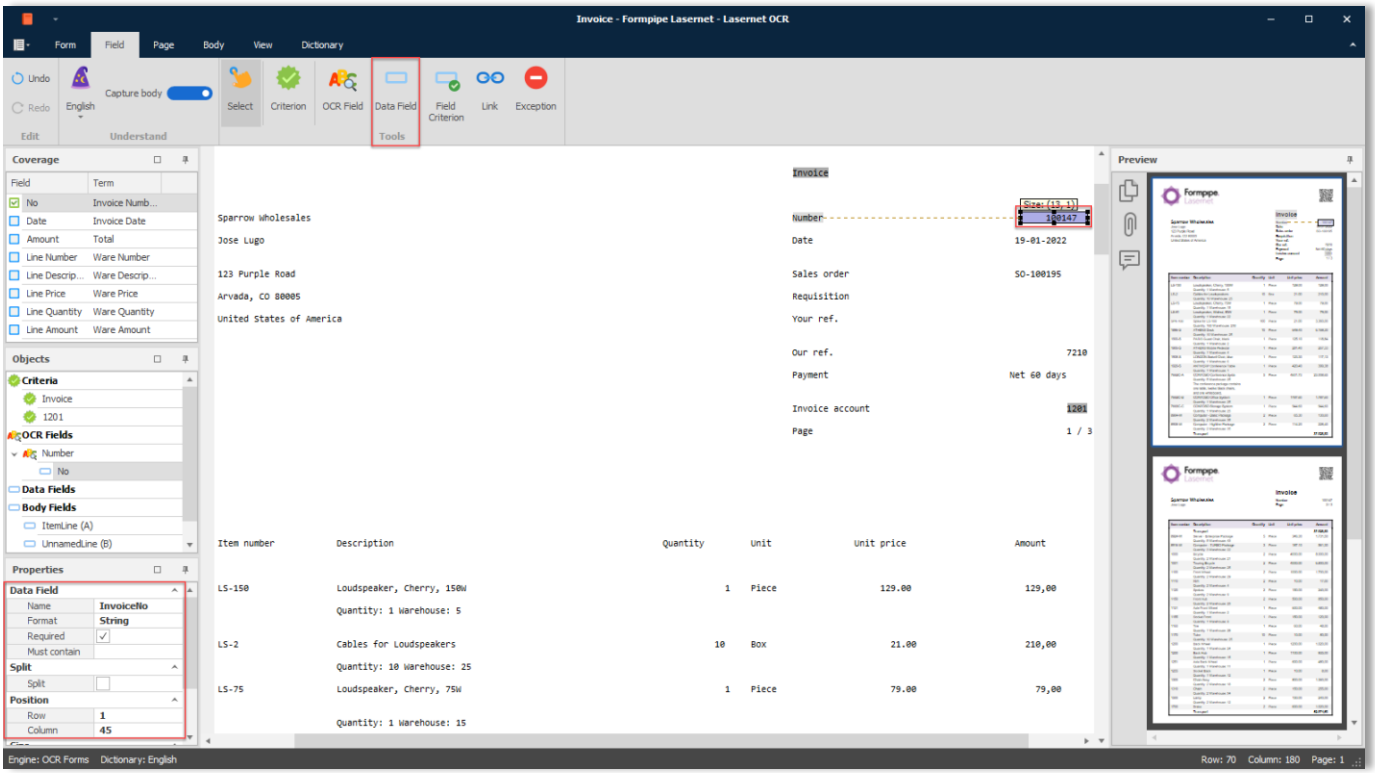
You can restrict your search to a user-defined area by adding values for column/row start/end, or you can use a mix of values and wildcards.



A yellow zone will surround the OCR Field in where the search text must appear. It is recommended that the size of the zone is larger than the text string for an improved search, especially for scanned documents.

Data Field

A Data Field is marked as a blue object: 100147



Name and Format are defined in the properties of a data field. If a data field is linked to an OCR field, the value of the row and column is in a relative position to the OCR field.

If the data field is not linked to an OCR field, the row and column values are in an absolute position to the start of the form data.

Objects List View

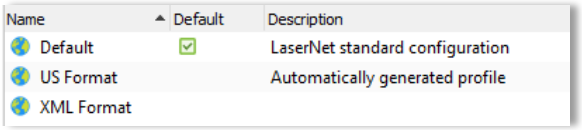
Data fields linked to an OCR field are listed as a child to a specific OCR field.

Data fields not linked to an OCR field are added to the list of general data fields.

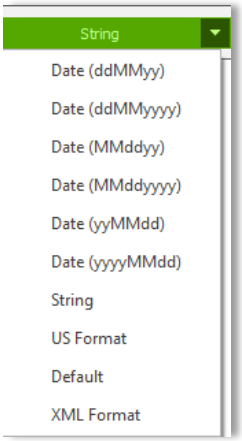
Name Name of the field. The list of available field names is obtained from the document type, managed in the OCR Engine.

The Field Name and its value will be included in the output format created by the OCR Engine (also viewable from the Form ► Output View).

Format Name of the Regional Profile. The list of available profiles is retrieved from the Lasernet Server on start-up. A field can either be a string, number or date (see Field Names list). Numbers and Amounts must match what is specified in the current profile (via the Regional Profiles), to successfully convert them into the final output format.



(An example of Regional Profiles created in the Lasernet Developer and accessible from the OCR Editor)



A range of predefined parameters for Dates and String are available.

The output result for a string field will be exactly as the data is presented in the document.

When using a Regional Profile for number and date formats, the separators specified in the Regional Profile must match the separators in the value of the field, otherwise the field will be ignored.

Required (1) Activated by default. This flags differently in the header/footer as opposed to the body.

Header/footer: If this setting is not active, the Field Name will not be represented in the output format.

Body: A valid value is required for the row to be included in output.

Text Used as an exact match of a Data Field. If a text string has been entered into this field only Data Fields containing this match will be captured. Wildcards are not supported.

Split If a Data Field contains two different kinds of data information, for example, a string and a number, or a number and a date, and there is no logical way of creating two Data Fields due to overlapping, the Split Field must be activated. This typically happens for two columns listed in the body area after a PDF or TIFF document is converted into Text.

Name (2) Name of the second field when running in split mode (works similarly to single Field Names).

The rule for splitting strings into two Field Names is that first Field Name contains the value of a date or a number, if represented in the string.

The second Field Name will contain the rest of the string.

If a number or date is not detected, both Field Names will be set as text strings. The splitter finds the first space character and adds the text string before it into the first Field Name and any character after the first space into the second Field name.

Format (2) Name of the Regional Profile for the second Data Field (works similarly to single Field Names).

Required (2) Activated by default. This flags differently in the header/footer as opposed to the body.

Header/footer: If this setting is not active the Field Name will not be represented in the output format.

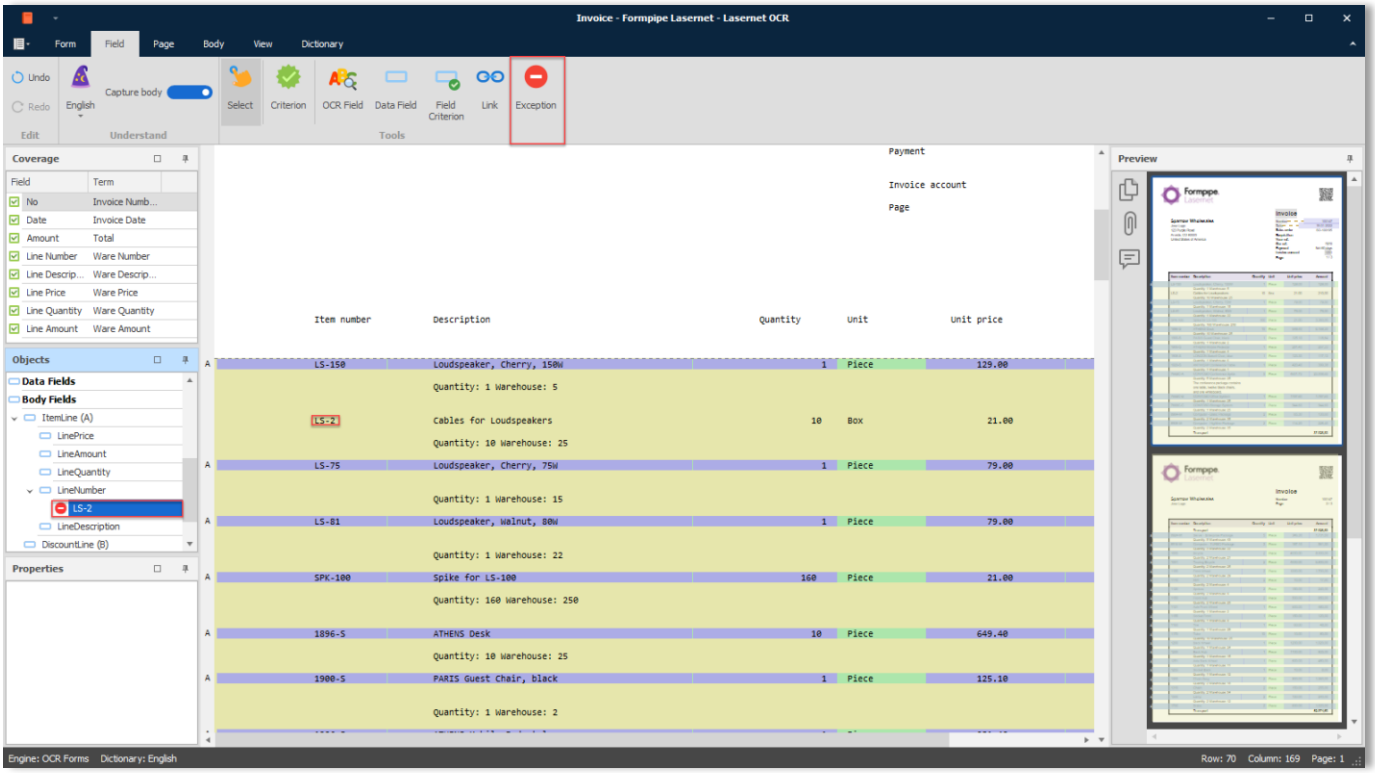
Body: A valid value is required for the row to be included in the output.

Row / Column Position for a Data Field represented as rows and columns in a text document. Position of a Data Field is relative if linked to an OCR Field. If not linked, the position is fixed.

Link No selections are available for the Link tool. The tool is used for linking a Data Field with an OCR Field; a dotted line shows which fields are connected.

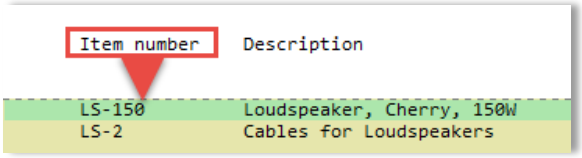
Exception No selections are available for the Exception tool. The tool is used in the body area as an exception for excluding patterns with specific text symbology. The

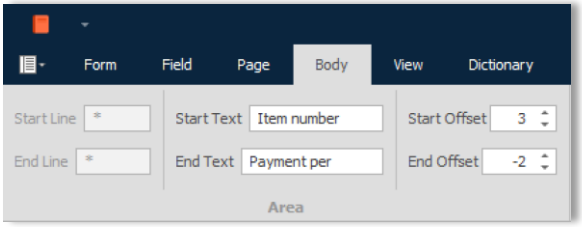
value of an Exception is connected to a Data Field and is shown in the list view along with active objects in the OCR Form.

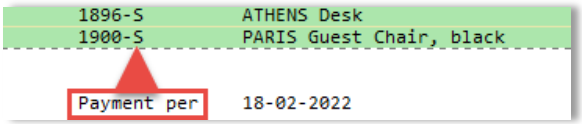


3.6.6 Body Properties

The OCR toolbar enables you to set up the body area so that a dynamic number of article lines can be detected and used in the final output format.



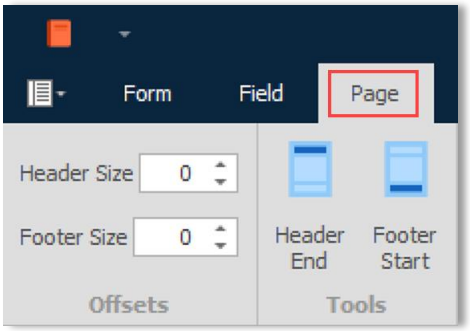




Start Line	Set fixed number of lines from where the body area starts within the input data file. The start line number must be a positive numeric value. This value is only required if a fixed start text cannot be found.
End Line	Set fixed number of lines where the body area ends in the input data file. The end number must be a positive numeric value higher than the start line number, or an asterisk (*) which means that the body stops at End of Page (as shown in the example). This value is only required if a fixed end text cannot be found.
Start / End Text	The body area can have its own Start and Stop text criteria which gives you flexibility beyond just defining its location in an input data file, by ensuring the data also satisfies the criteria set.
Start / End Offset	Used together with start criteria in the body area. The header (start) or footer (end) area can either be extended with a positive number of additional lines or decreased with a negative number of lines.

3.6.7 Page Properties

Each page will typically contain a header, body and footer. Very often the numbers of printed text lines, for the header and footer size, are defined with exactly the same size. In the Page tab you can define the size of the Header and Footer of the OCR Form.

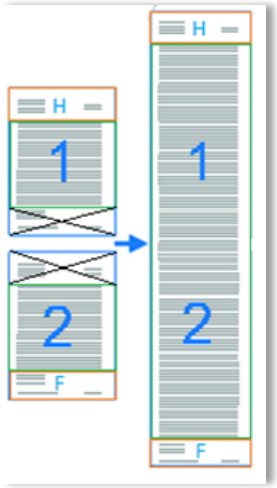


We recommend these settings are carefully considered, because important data can be lost if the defined size does not fit the form's conditions. The benefit of removing headers and footers from middle pages in a document is easier maintenance of item lines.

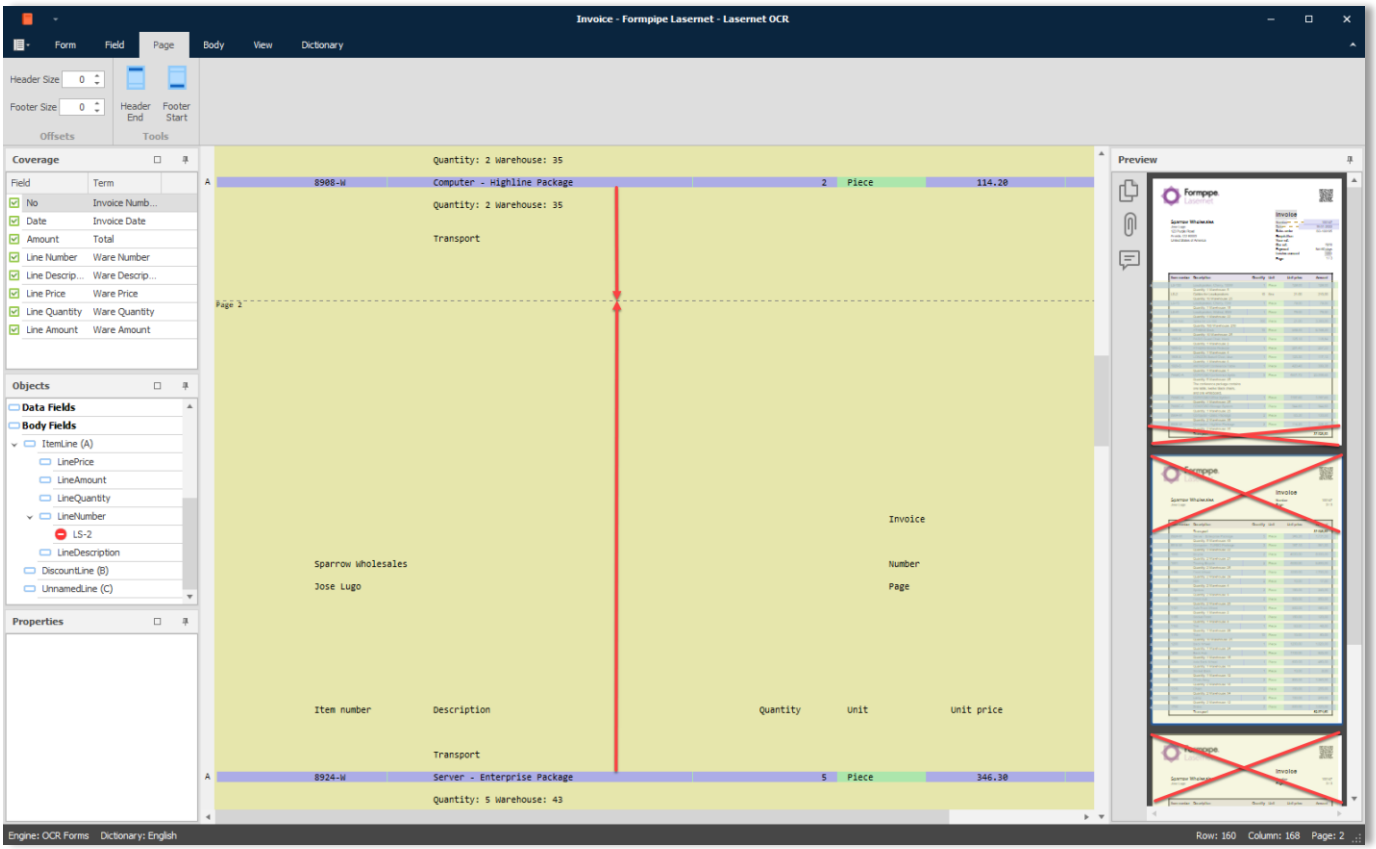
Header Size	The number of lines for header size; if the number of lines in the header is exactly the same for all pages you want to combine.
Footer Size	The number of lines for footer size; if the number of lines in the footer is exactly the same for all pages you want to combine.

The effect will be that the following will be combined / converted into one grab file / job, in the illustration:

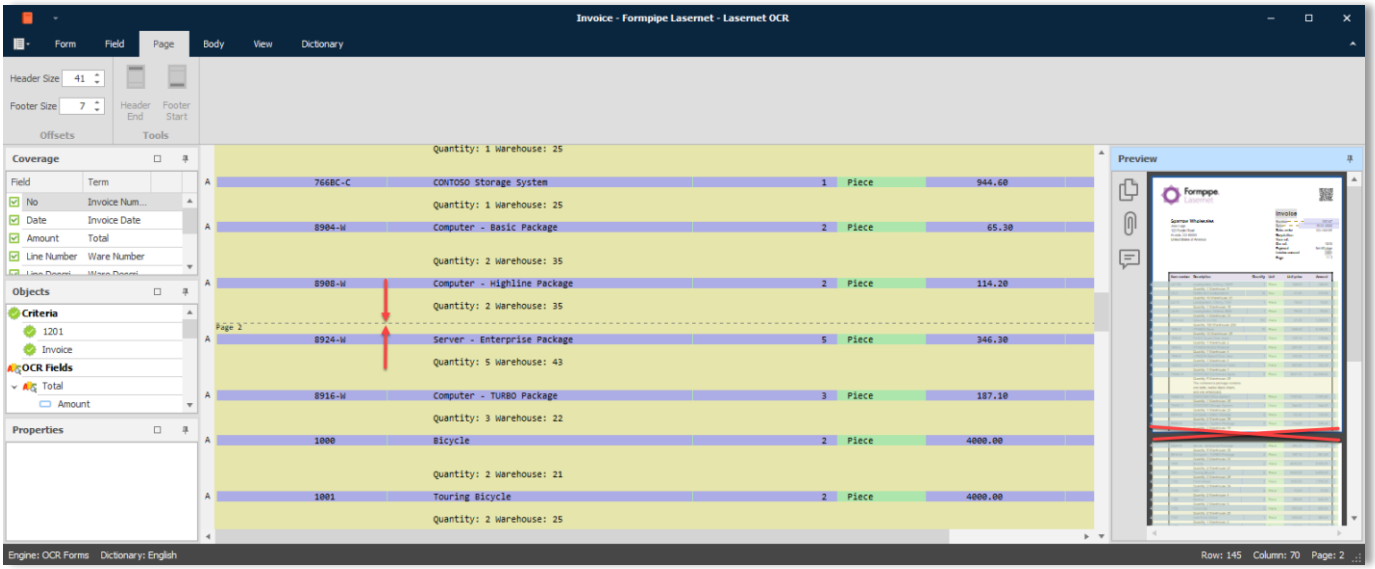
- the header for the first incoming page
- the body lines for all incoming pages
- the footer for last incoming page



By default, the headers and footers for all pages will be visible in the data. This is not critical because the item lines are recognized as expected and the headers and footers in between are ignored.



In this example, the Header Size is 29 lines and Footer Size is 9 lines. The pages are combined and only one header (first page) and one footer (last page) appear in data.

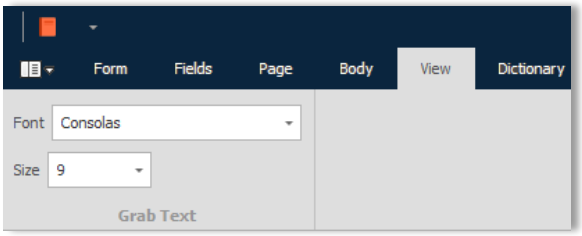


3.6.8 View Properties

Set up the preferred font type for grab text.

Font Font character name. The default is Consolas.

Size Font character size. The default is 9.



Note: None of the settings will have effect on the final output. This is for viewing only.

3.7 Walk-through – How to create an OCR Form

3.7.1 Create a Form

When you have added an OCR Engine to your project and a workflow to convert incoming PDF or TIFF files to a Text file, you are ready to start working with OCR Editor.

OCR Forms cannot be designed without having some current (or sample) business data to work from. This data needs to be stored in a text file on a share that is accessible to the OCR Editor, or saved in paused mode in the internal Lasernet database.

You can create a new OCR form by following these steps:

- **Scenario 1** – Lasernet has converted and saved a text copy in a shared location.

  Start OCR Editor. When the application is open go to File ► Grab file and browse for a valid text file.

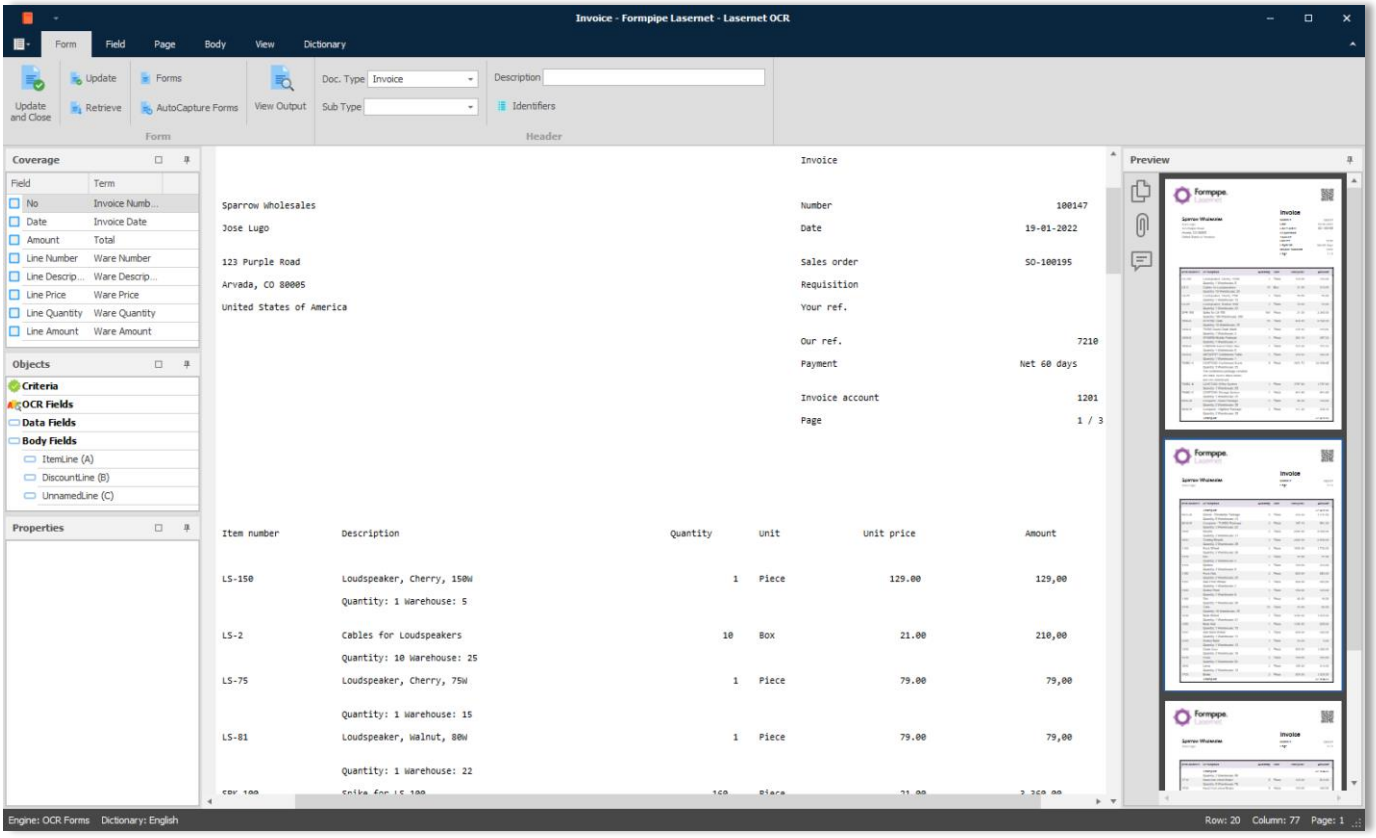
- **Scenario 2** – Lasernet has converted and saved a text copy in the internal database.

  Start Lasernet Client. When the application is open, click the Paused tab and double-click on any of the documents listed for the OCR Engine. The OCR Editor will open the document.

In this guide we will set up an incoming invoice from a supplier named “The Factory”.

A PDF document is received by email and converted into a text format and paused by the Lasernet Server.

The document is opened in the OCR Editor to map OCR fields and extract data for a workflow system.

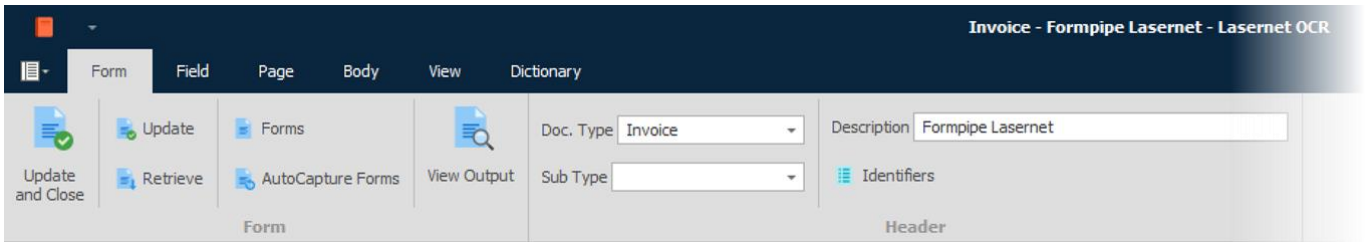


The PDF file will contain logos and graphical elements, but they are ignored and not needed to map the OCR fields. The key data (text) will be inserted in to rows and columns in a text file. The preview window will show a copy of the original PDF document.

ℹ When the text / grab file is loaded in OCR Editor, you are ready to create an OCR Form.

3.7.2 Setting up Form Properties

Set up the form properties. Click the Form tab.



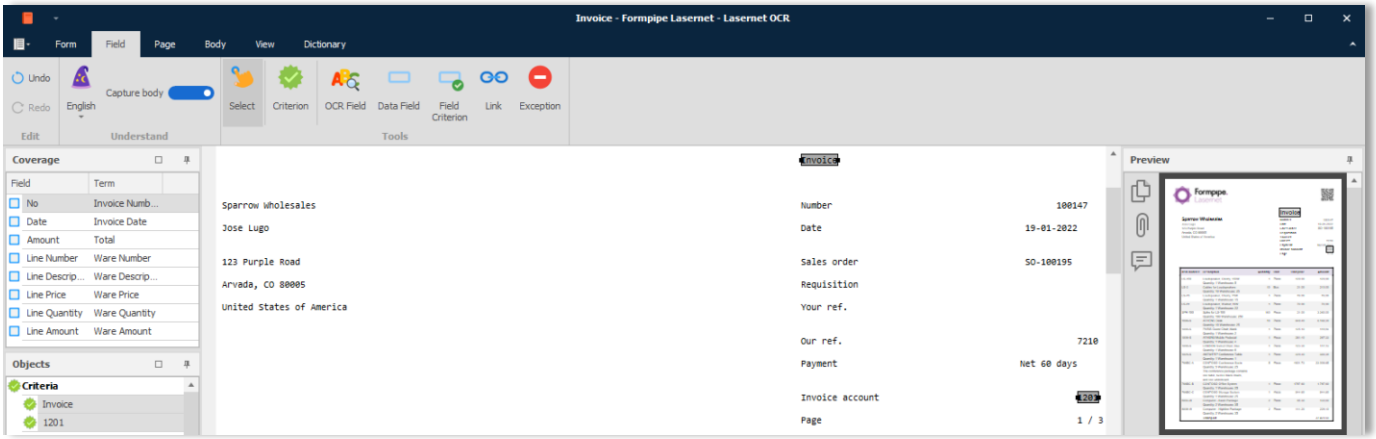
From the Doc. Type drop-down menu, select the proper document type. In this example, we have selected **Invoice** as document type and that the contents belong to the company called **Formpipe Lasernet**. A list of required OCR fields is defined in the OCR Engine for the invoice and they will now appear in the Coverage window as a kind of to do list.

3.7.3 Setting up Criteria

To successfully match an invoice from “Formpipe Lasernet”, we must set up a couple of unique criteria. In this example, we are setting up two criteria. One matching the document type “Invoice” and another matching the Invoice account “1201”.

Click **Criterion** and mark the form name “Invoice” by drawing a box in the text string. Then draw a second box around value “1201” for the invoice account.

Note: A criterion match can only match a single line of text.



Two dark grey objects show how to match the OCR Form:

Documents that are processed through the OCR Engine and match the two strings will be identified as an **Invoice** from invoice account **1201**.

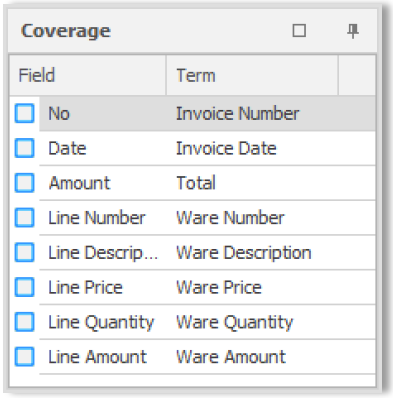
Add additional criteria to provide greater differentiation between other processed documents. Matching only the word 'Invoice' would match any invoice from any supplier.

3.7.4 OCR Field

The next step is to set up the OCR and Data Fields. Select the **Field** band.

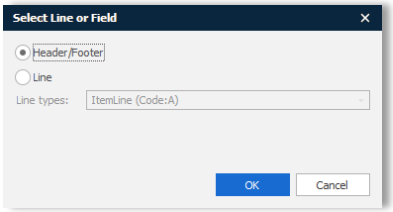
Firstly, you need to pick up the invoice number.

Select the **InvoiceNo** in the **Coverage** list:



A wizard will guide you through the required selections for marking / selecting the OCR Field, Data Field, Field Name and Field Format.

Mark the first OCR Field, containing the Number, by drawing a box around the text string or double-clicking on the word.



Select whether the Number exists in the Header/Footer or is a part of an item line. In this example, the label Number is positioned in the header. Click **OK** to confirm.

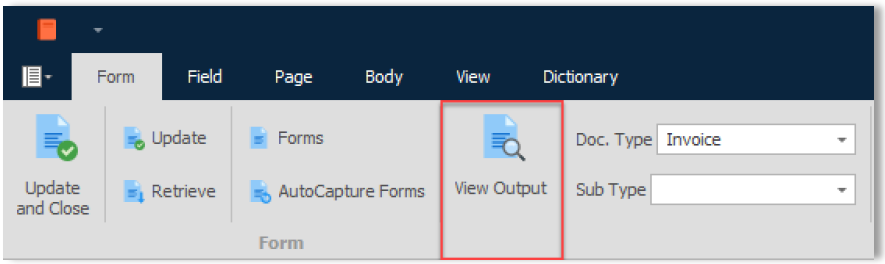
To mark the actual invoice number, click the Data Field and draw a box around the invoice number 100147.

When the Field Name is selected, the wizard will automatically close.

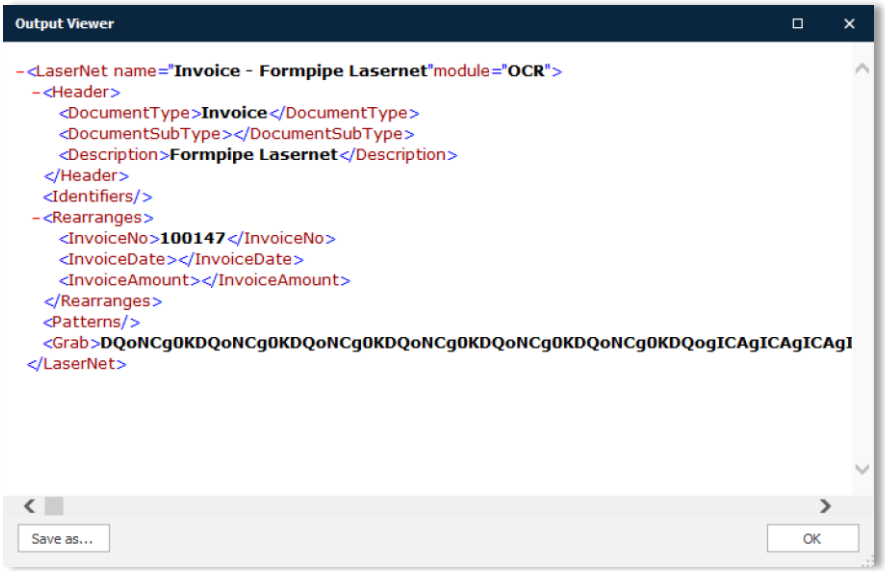
Note: The options in the wizard vary depending on the selected field types because a field format is required for date types and amounts.

Two new objects are created. The light grey one is the OCR Field, and the blue one is the Data Field. A dotted line between the two objects shows that they are linked together and that the position of the Data Field is relative to the position of the text string “Invoice No.”.

The first field is mapped for The Factory. If you want to view the result of the extracted data, select the Form tab and click **View Output**.



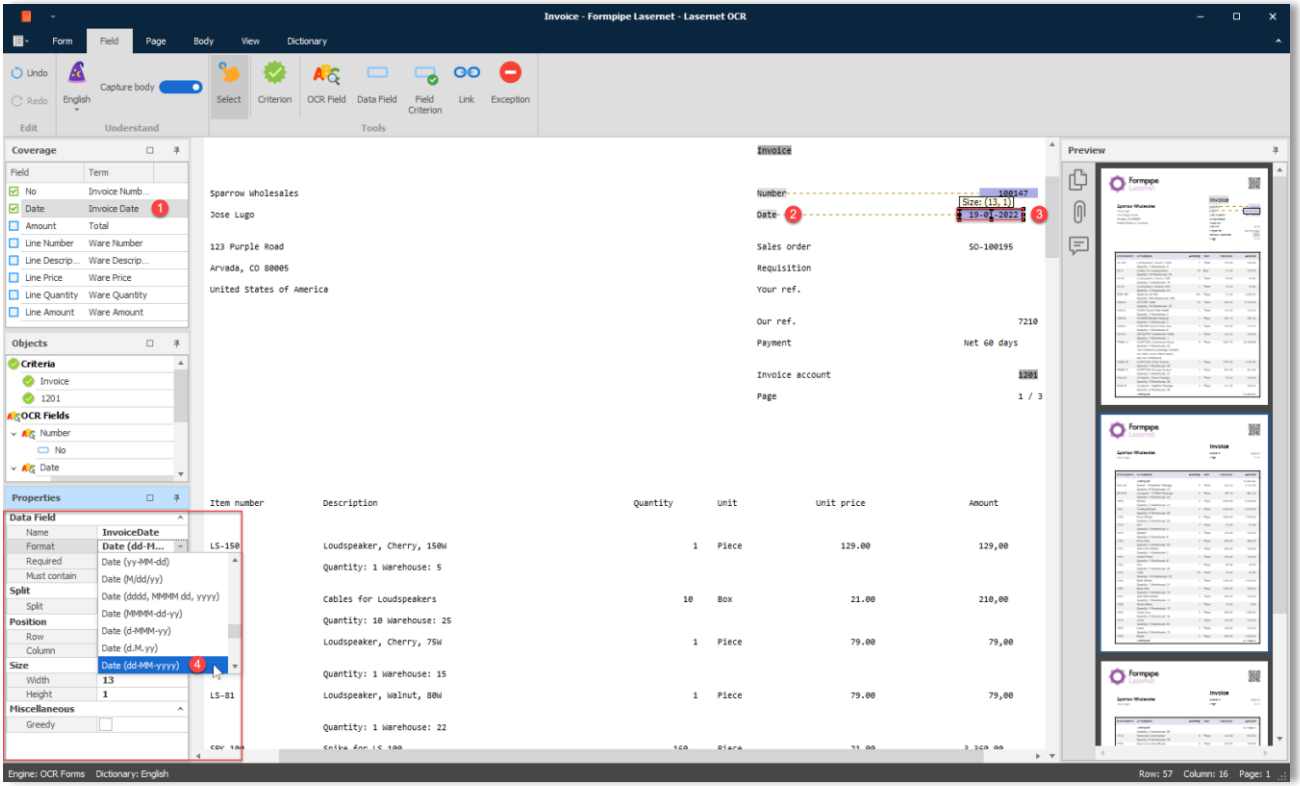
This is the preview result of the pre-defined XML format created by the OCR Editor.



The next step is to go through the other fields from the coverage list.

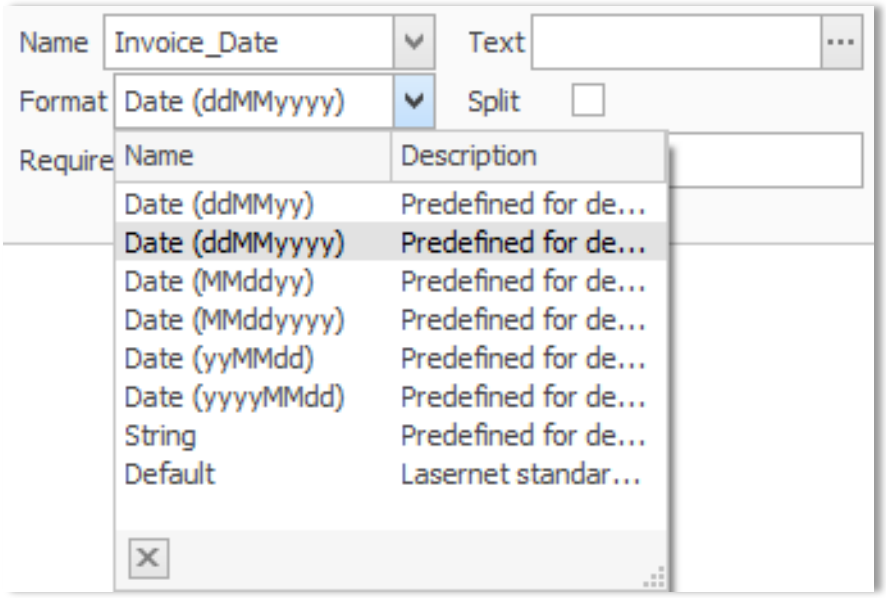
We will map the invoice date by starting the wizard from the Coverage window.

Click on the InvoiceDate field in the Coverage window and follow the same steps as in the previous example. Select text string “Date” and select InvoiceDate as Field Name. Mark “19-01-2022” as the value for the Data Field.

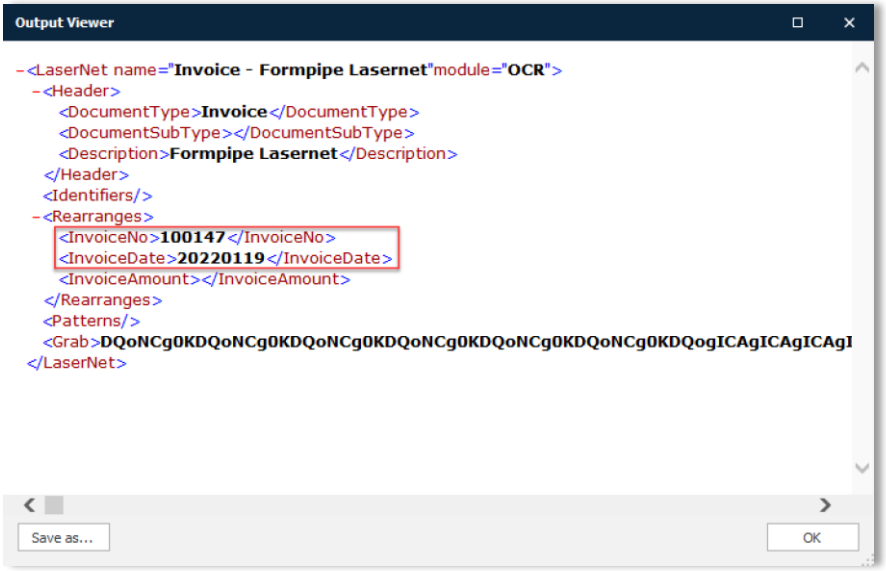


Two new objects are created; an OCR Field recognizing the Date string and a Data Field to capture the date value.

Set the proper data format for the Date field. In this example, the format must be set to **ddMMyyyy** to match **19-01-2022**.



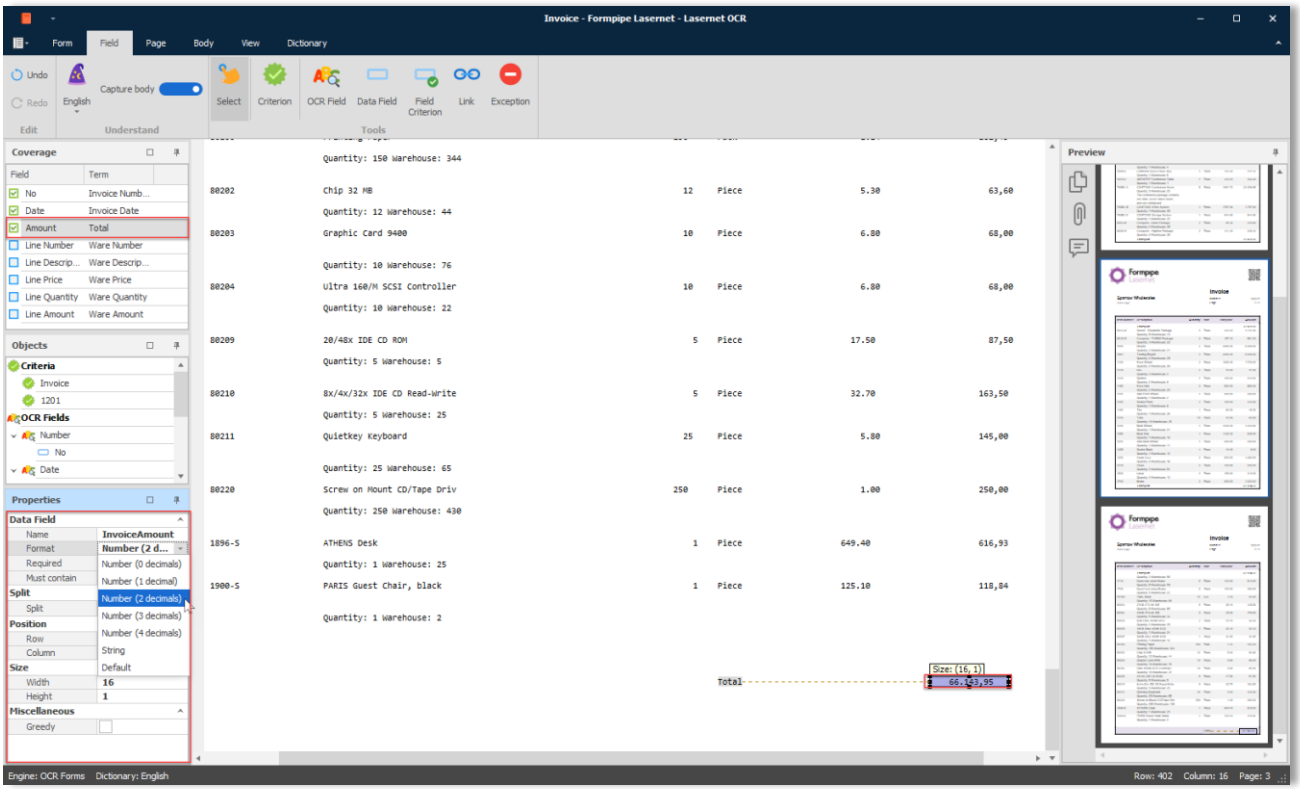
Two fields are now mapped. Go to the Form tab and click **View Output**.



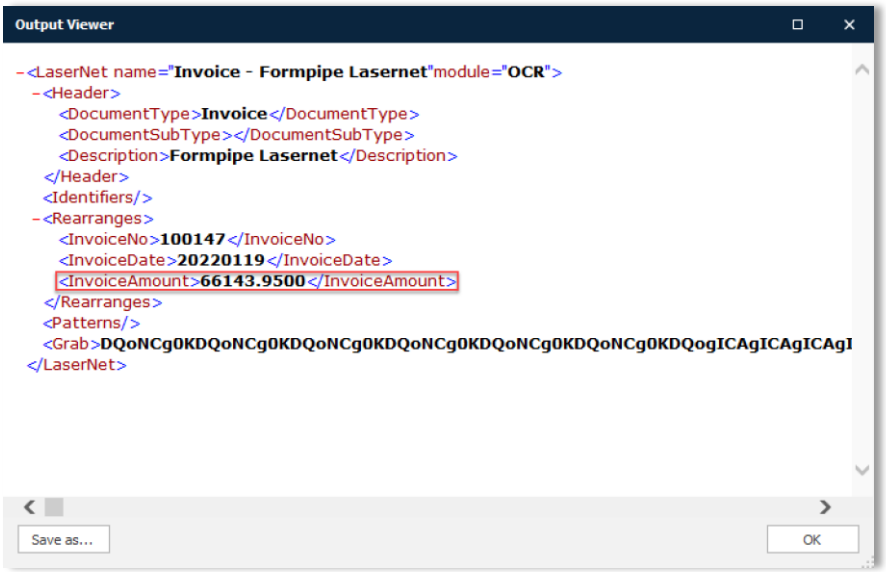
The date value is now added to the XML structure and converted to a standard format yyyyMMdd**,** which is the preferred date format for most workflow systems.

Finally, we will set up OCR and Data Fields to match the total amount of the invoice. The amount will be captured in a position relative to the text string “Total”.

Select the TotalAmount in the Coverage window and go to the wizard to create the OCR Field and Data Field. Set Field Format to a Number with 2 decimals.



We have now added the total amount to the list of covered OCR Fields. Go to the Form tab and click **View Output**.

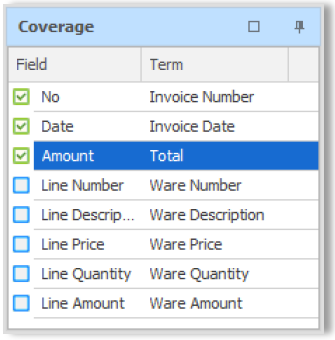


The value for the total amount is added to the XML structure as a standard number format with four decimals.

You are now ready to add the rest of the required fields to the OCR Form.

3.7.5 Covered Data Fields

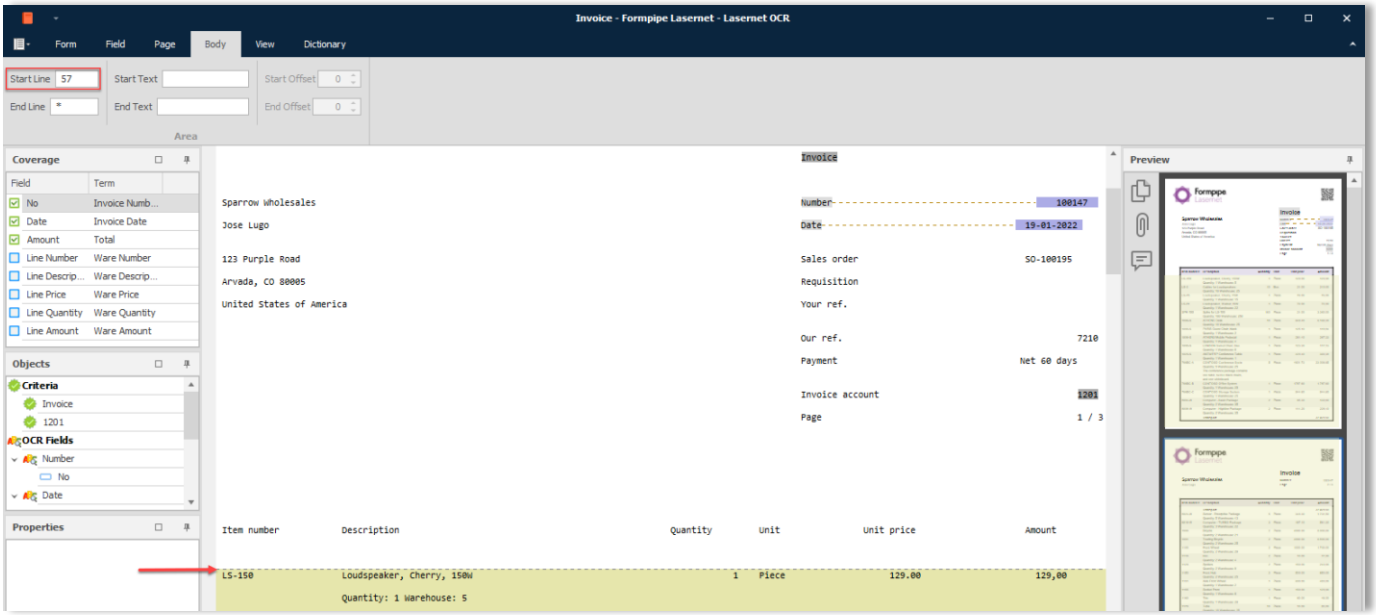
In the Coverage window, a check mark indicates the Data Fields that are covered in the form. This is a guideline to help you to cover all the required fields. The OCR Engine is still able to process OCR Forms even if not all of the fields are covered.



3.7.6 Creating the Body Area

The next step is to define the body area of the form to recognize the item lines and capture the data fields. OCR fields are not required in the body area; more complex algorithms are used to intelligently match the data fields instead.

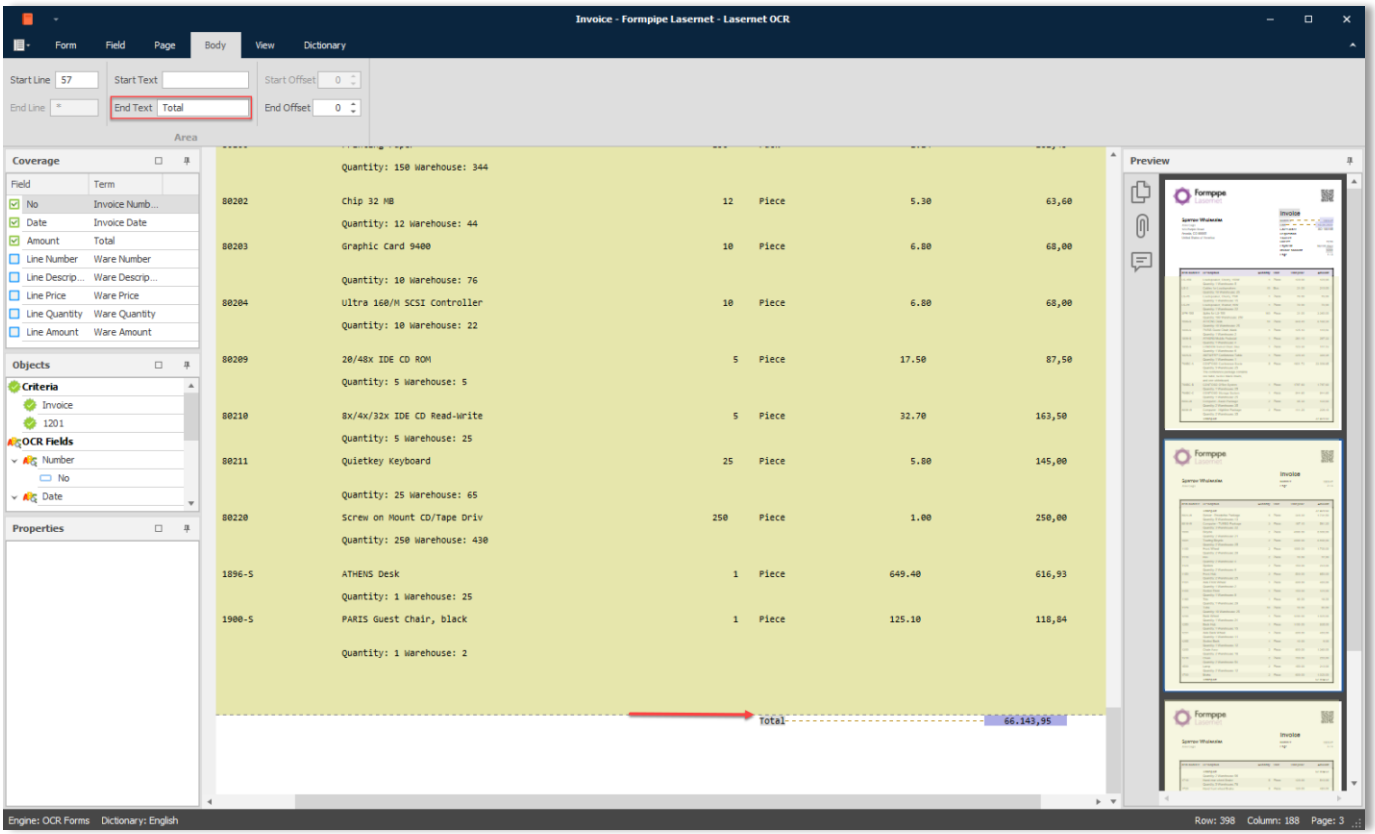
Click the Body tab and set Start Line to 36, to define the start of the item lines.



The end of the body can be set to either of the following:

- a fixed End Line value
- an End Text that marks the end of the body

In this example we will add value for the End Text string to define a body with a variable length. Type “Total” to locate the end of the body.



Additional settings to use an Offset relative to a Start Text or End Text value are also available. The values can be set to either a positive or negative value. An offset will include fewer or more lines in the body area relative to where the Start or End Text is located.

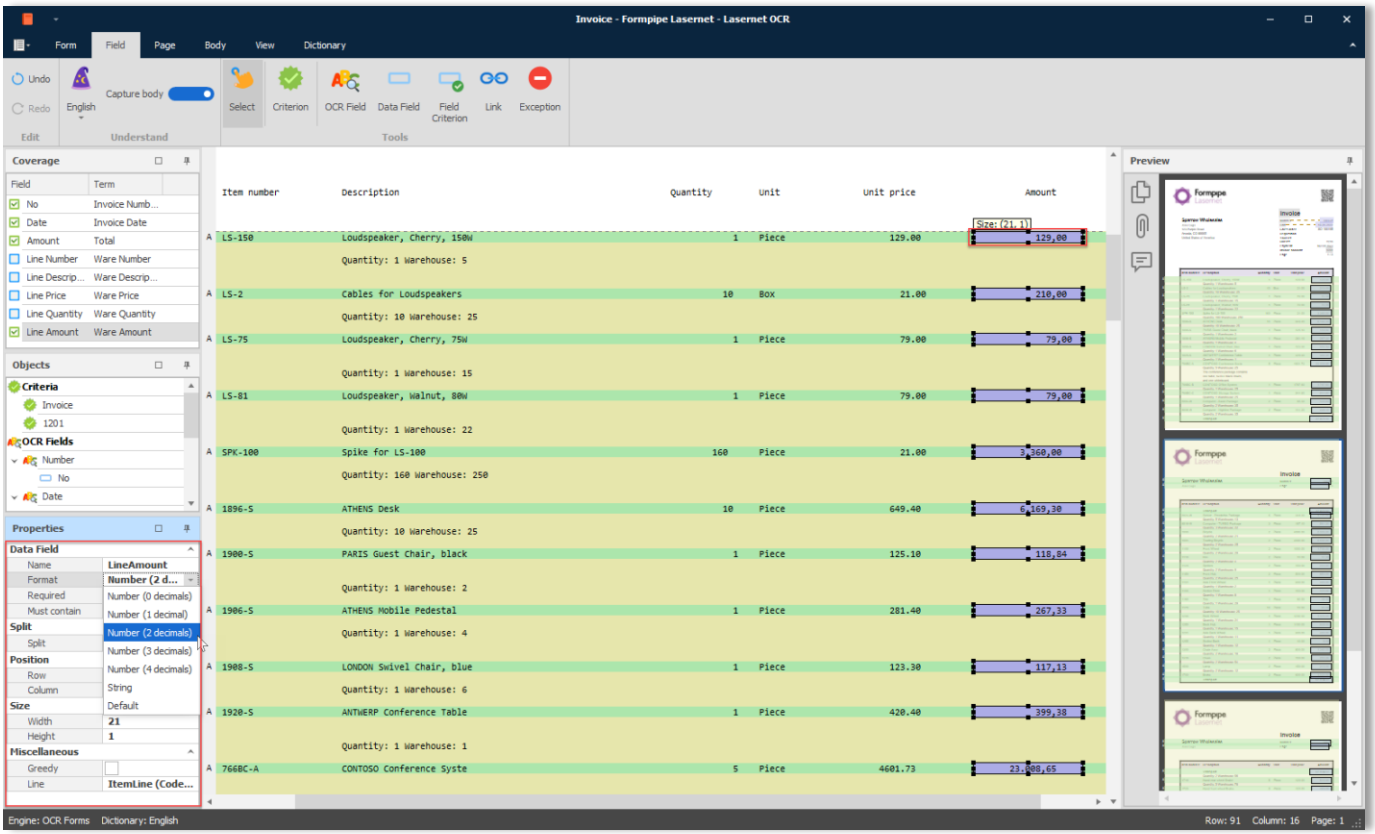
ℹ️ For more information about Start/End Offsets, refer to section 3.6.6 “Body Properties”.

3.7.7 Data Fields in Body

Next step is to cover the Data Fields that appear in the body.

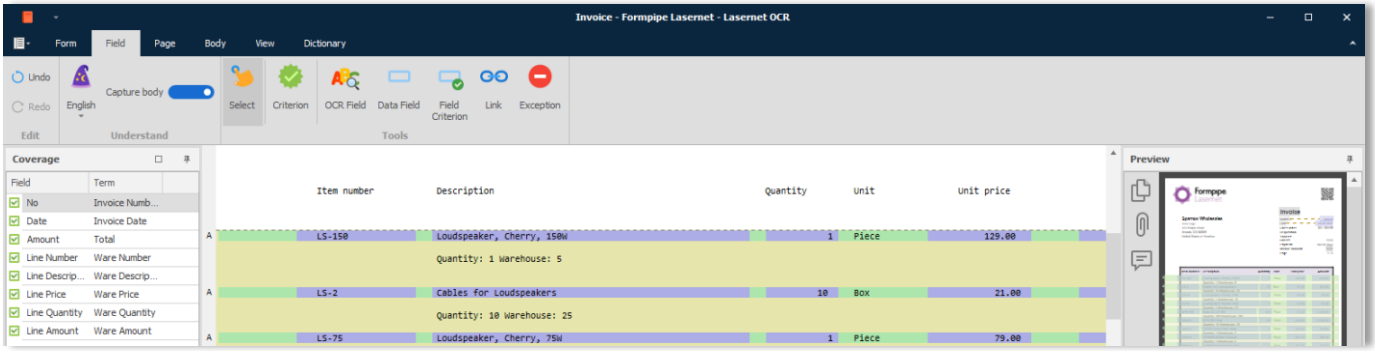
Select the Data Field tool or the Data Field name for Line_Amount from the Coverage window.

We recommend that you start capturing columns that contain numbers with decimals. The algorithms filtering the Data Fields and finding item lines are optimized for this purpose.



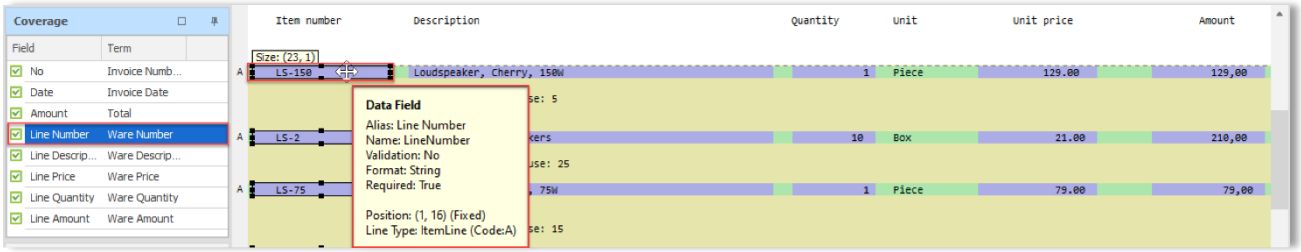
Because you do not need to create an OCR Field, you can select **Mark or select Data Field** in the wizard and then mark the column with the line amount. This type of Data Field is always located in a fixed column.

Draw a box around the value in the column and ensure that it is wide enough to contain the maximum number of characters possible for that column, otherwise data will get cut off or included in the wrong Data Field for other documents of same type.

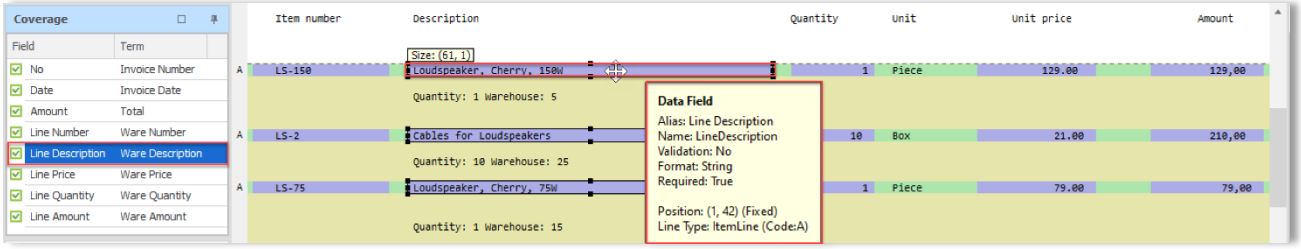


Next we cover the rest of the data fields from the coverage list, which are data fields with a mix of strings and numbers.

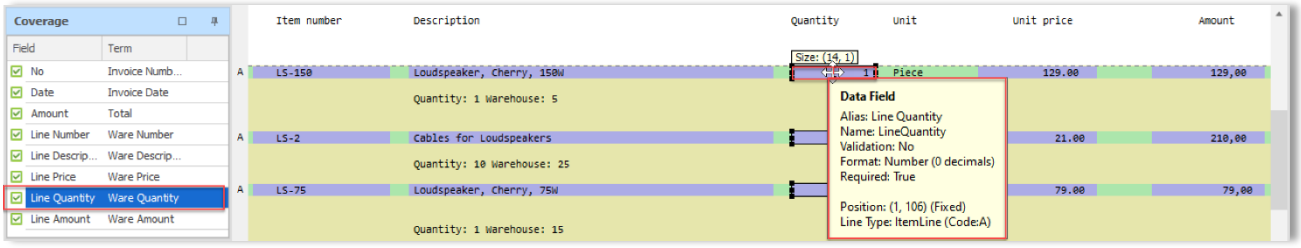
Name LineNumber
Format String
Required True
Size 23 characters



Name LineDescription
Format String
Required True
Size 61 characters

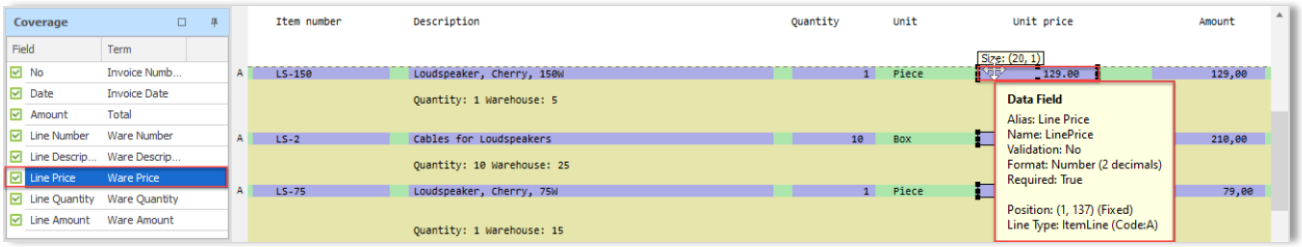


Name LineQuantity
Format Number (0 decimals)
Required True
Size 14 characters

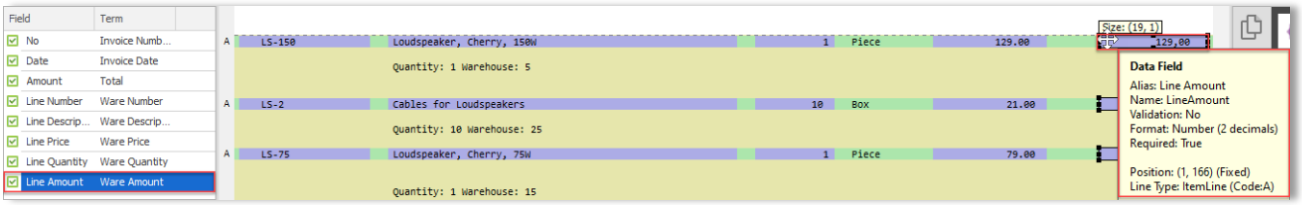


If you are sure about number formats, you can always set the format type to string. The value will then keep the same format in output data.

Name LinePrice
Format Number (2 decimals)
Required True
Size 20 characters



Name LineAmount
Format Number (2 decimals)
Required True
Size 19 characters

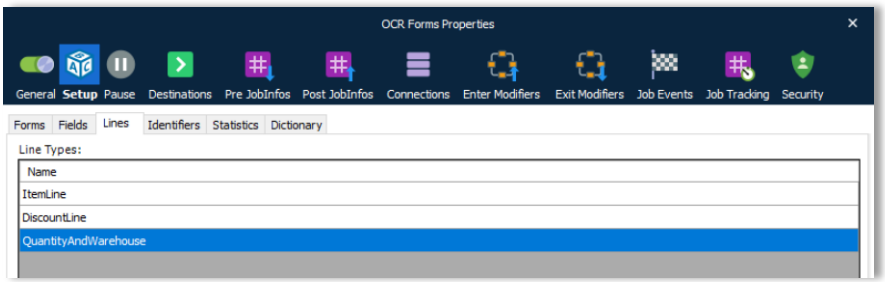


All the required fields are now mapped for the invoice from “Formpipe Lasernet” and the form is ready to be updated to the server.

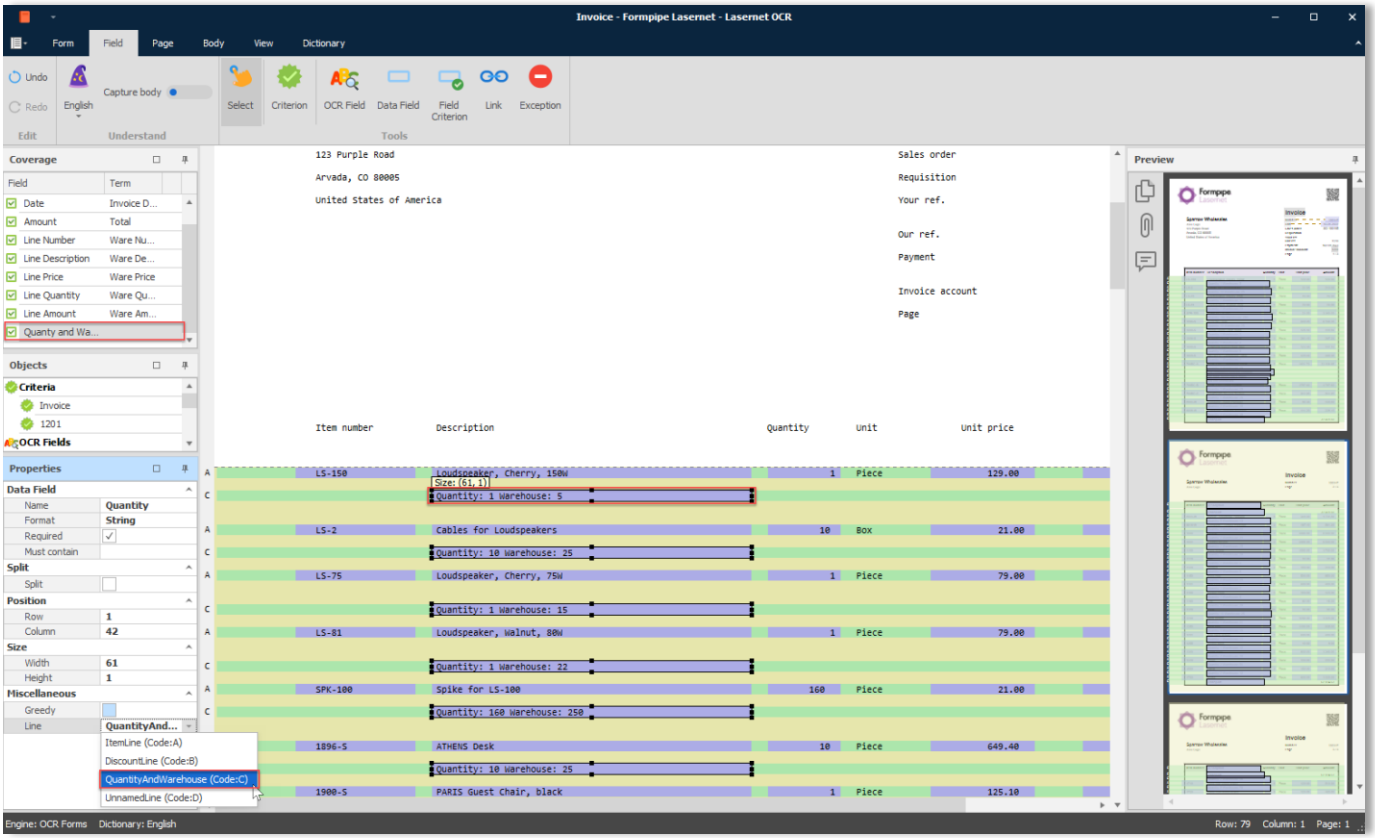
3.7.8 Additional Item Lines

Each item line with data you want to capture must be attached to a Line Type. It can be a line type configured in the OCR Engine on the Lasernet Server or you can assign a Data Field to **UnnamedLine**, which is a dummy line type always present in the OCR Editor.

If you have additional item lines, like a Quantity and Warehouse figures, you want to extract separately, you must have also an additional Data Field assigned to the line type QuantityAndWarehouse. Additional line types like this can be defined in the OCR Engine, on the Lasernet Server and selected afterwards in the OCR forms. In the XML output format, each type of item lines will be grouped individually.



Here we have an example of four types of item lines, configured in the OCR Engine, including the system defined UnnamedLine, that is always present in the OCR Editor.

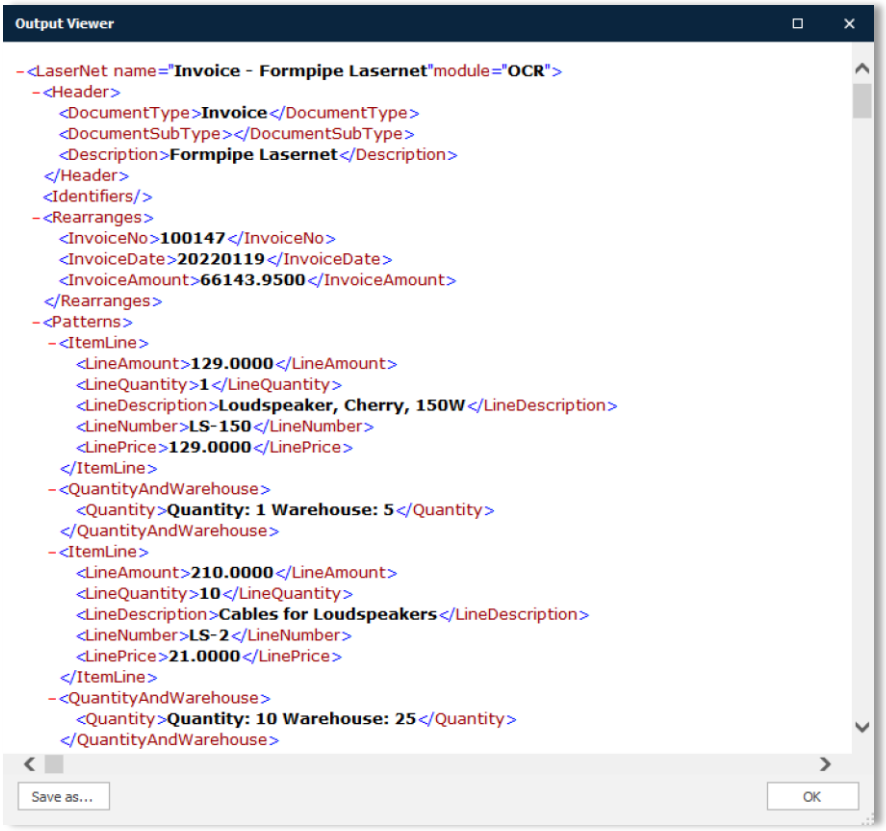


You must always find matches in the item lines to differentiate between individual type of item lines.

In the user interface you can see the difference between Line Types. In front of each line, you will find a letter (starting from A, B, C…). The letter will refer to which line type the pattern belongs.

3.7.9 View output

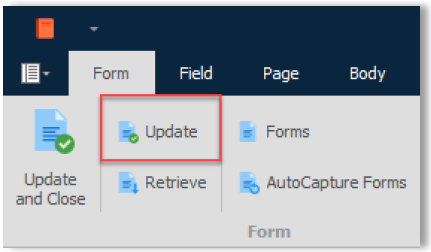
Go to the Form tab and click **View Output**.



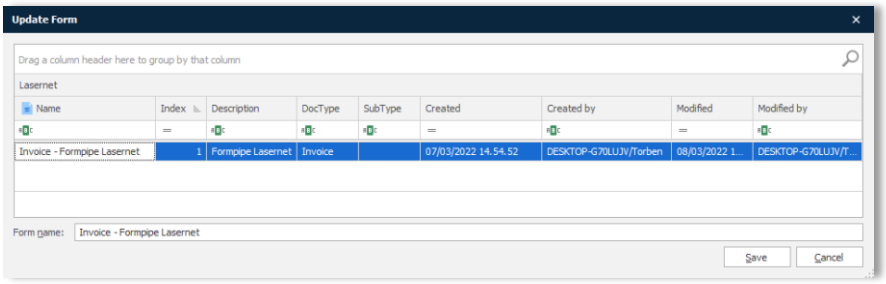
This is an example of the XML format created by the OCR Engine. All the mapped field names and values are included in the job.

3.7.10 Update OCR Form

When all the required fields are added to the OCR Form, you are ready to deploy the OCR Form to the Lasernet Server.



Select the Form tab and click **Update**.

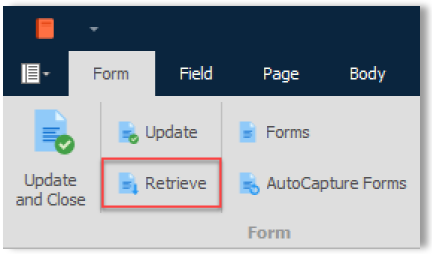


Type in a logical name for your OCR Form, usually the name of the invoice supplier - then click **Save**.

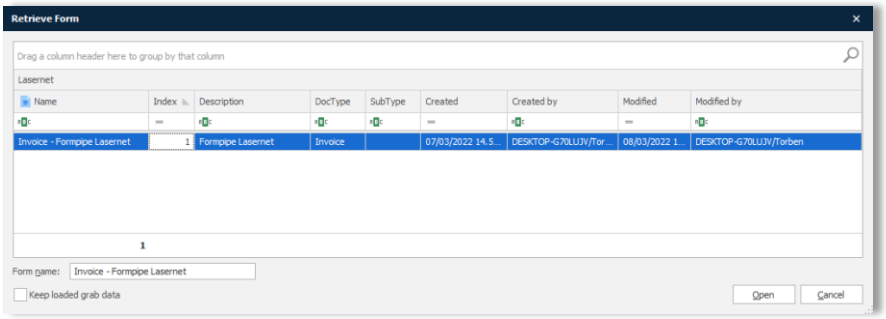
The OCR Form will be uploaded to the Lasernet server and activated. Documents received by the Lasernet server that match this document type and supplier name will be converted to an XML format that can be used by other modules in Lasernet or an external workflow system.

3.7.11 Retrieve OCR Form

You can retrieve and edit an existing OCR Form from the Lasernet Server.

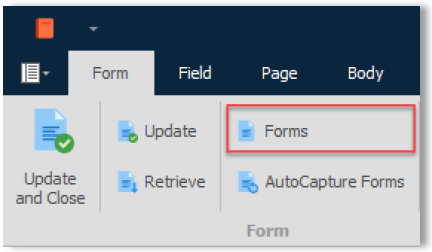


Select the Form tab and click **Retrieve**. Changes can now be made to the OCR Form and then be uploaded, with the new changes, to the Lasernet Server.

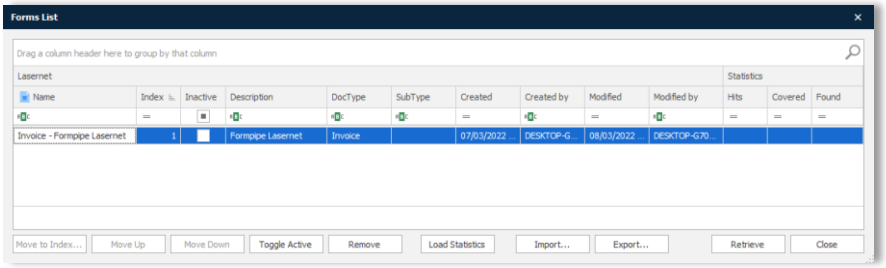


3.7.12 Forms List

Select the **Forms List** button for the following actions: view **Created/Modified By**, **Move to Index**, **Remove** discontinued OCR Forms and **Import/Export** OCR Forms between Lasernet projects.

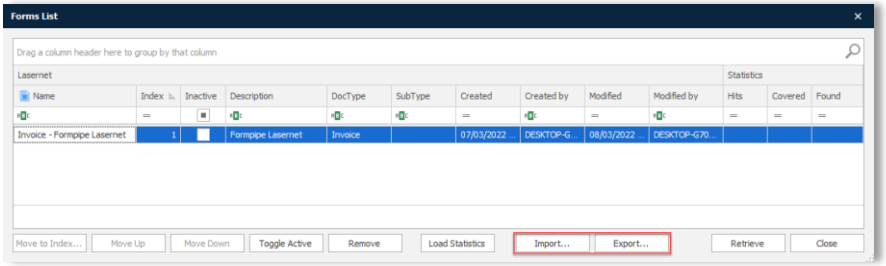


We recommend that you move regularly used OCR Forms to the lowest index number for the best performance. Documents are analyzed in index order and the first match for a form criterion will extract the OCR Fields based on the defined rules in the recognized form.



Mass input/export support

Select the Forms List dialog to Import or Export a single or multiple OCR Forms. Mark the OCR Forms in the grid and click **Export**. Select a temporary folder to which you want to export.



OCR Forms can be imported to other Lasernet projects. Open the OCR Editor application for another project and select the Import tool.

Browse to the folder to which the OCR Forms are exported. All the OCR Forms located in the selected folder will be imported. You must manually remove the files from the folder, before the import action, to exclude specific OCR Forms from being imported.

If the name of an imported OCR Form already exists, a new copy of the OCR Form is created. The new copy will be appended with an "_2" suffix.

During import, the OCR Forms will be immediately deployed to the Lasernet server. It is not necessary to manually run the Update functionality afterwards.

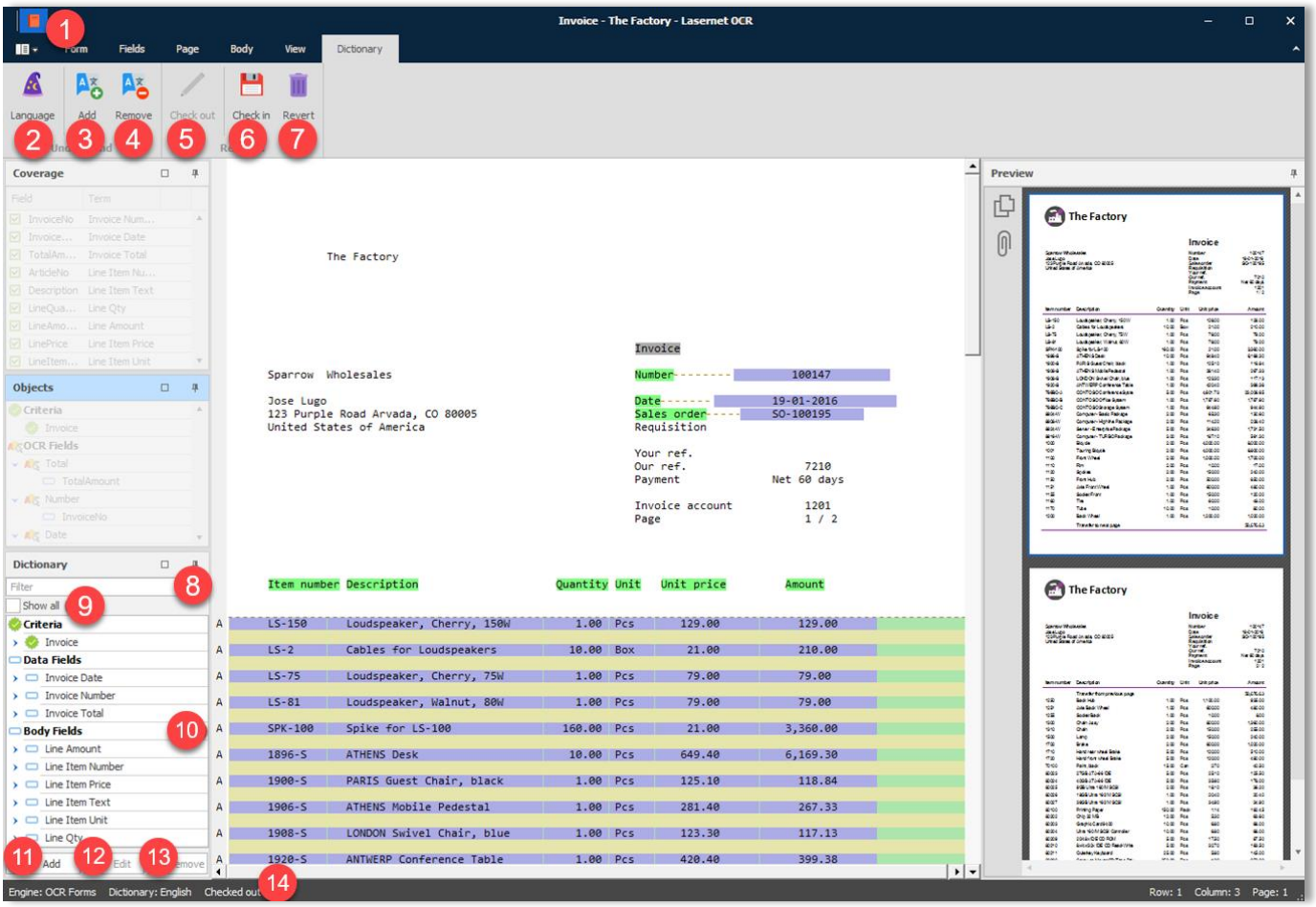
3.8 Dictionary Service

In the previous chapters we covered how to create OCR Fields with the built-in OCR tools and wizard to create Form Criteria, OCR Fields and Data Fields.

The solution also provides a Dictionary tool to automatically create objects with known phrases in the OCR Form. This requires a central database with a list of terms and regular expressions to recognize labels and extract data from documents.

3.8.1 Manage Dictionary

In the OCR Editor you can manage the Dictionary by clicking the Dictionary icon (1) in the application. In this mode you can only manage the Dictionary. All settings and functions used to manage OCR Forms are disabled.



Select which Language you want to manage words for in the Dictionary. Before you can add or manage entries to the selected language, you must **Check out** (5) a private copy of the current version. Click **Check in** (6) to make your changes available to other users of the Dictionary. The selected Language and Check out status are available in the Status Bar (14).

Tool bar

Language Select which **Language** (2) you want to manage words for in the Dictionary.

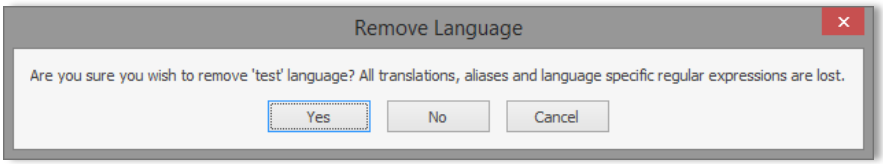
Add

Enabled in Check out mode. Click **Add** (3) to add a new Language to the Dictionary.

Remove

Enabled in Check out mode. Click **Remove** (4) to remove a Language and all contents in the Dictionary for the selected Language.

A warning message will be shown stating before all translations, aliases and language specific regular expressions are lost.



Check out

Click **Check out** (5) to prepare a new revision of the Dictionary. A private copy of the Dictionary will be created. Changes made to the Dictionary in Check out mode will not be available for other users after a Check in.

Check in

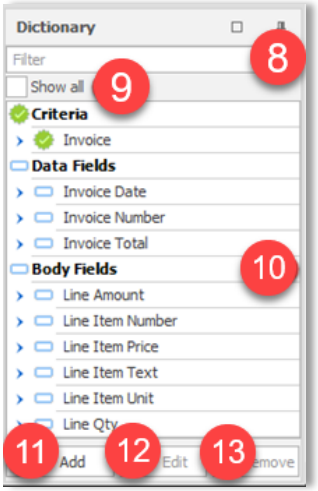
Enabled in Check out mode. Click **Check in** (6) to update the Dictionary and deploy the modifications to other users of the dictionary.

Revert

Enabled in Check out mode. Click **Revert** (7) to discard all changes made to the Dictionary since last Check in.

Dictionary

You can View, Add, Edit and Remove Terms from the Dictionary in the tree view in the Dictionary pane.



Filter

Type a string to **Filter** (8) the contents of the Dictionary by.

Show all

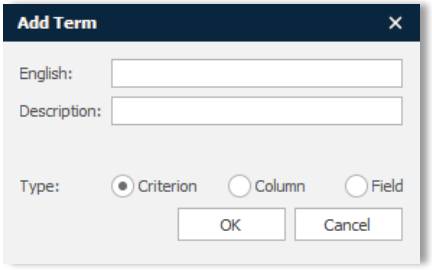
Enable **Show all** (9) to view all existing entries in the Dictionary for the selected language. Disable to show only the entries in the Dictionary that exist in the OCR Form that you are currently working on.

Dictionary

The list of Criteria, Data Fields and Body Fields available in the **Dictionary** (10) for the selected Language.

Add

Click **Add** (11) to manually add a new term to the Dictionary. Type the name in **English** for the term and a **Description** to describe the usage of the term.



If you have selected a language other than English, an additional field for the selected language will be visible. In the field you can type in a translation for the language you are currently working with.

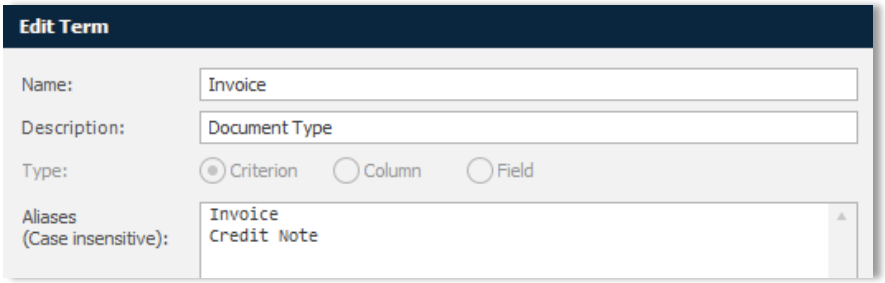
In dialogs, where the Term is present, the translated Term will be shown. The Term is bound to its English name and this will be the primary key in the Dictionary.

Edit

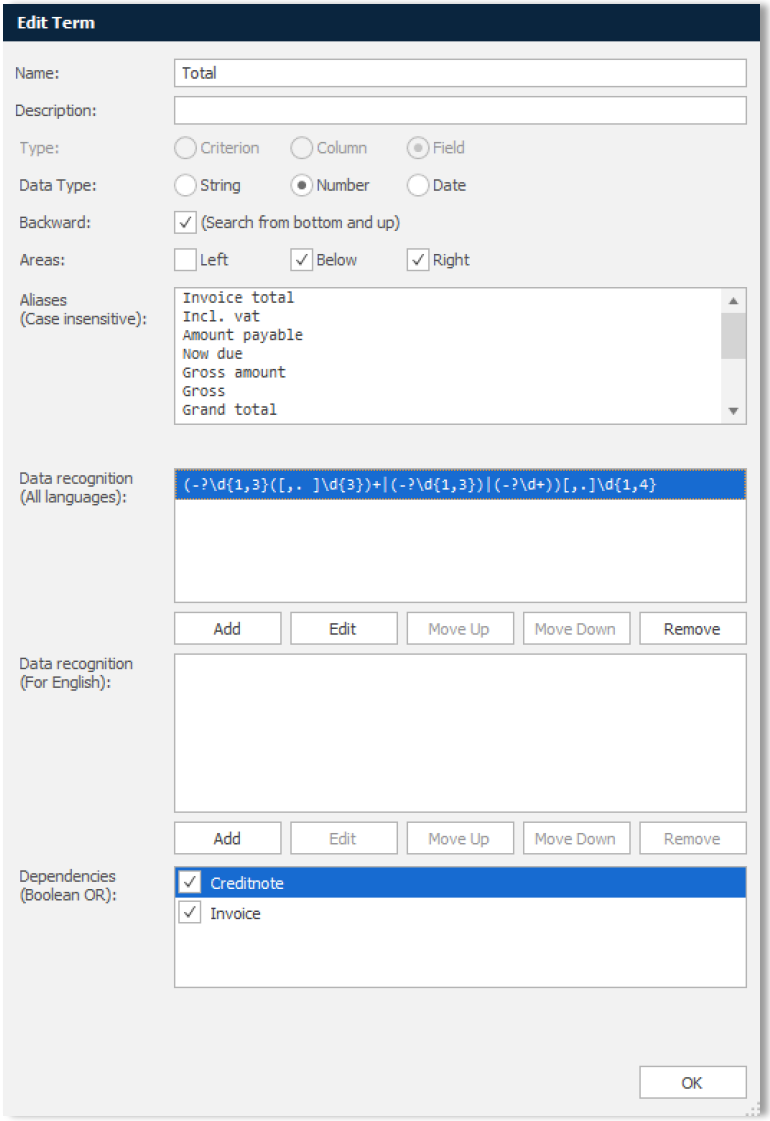
Select a Criterion, a Data Field or a Body Field in the Dictionary. Click **Edit** (10) to change the contents of an existing Term.

Note: There are two variants of this dialog, depending on the context.

Edit Term – Criterion

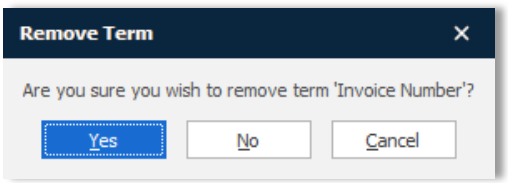


Edit term – Data Field



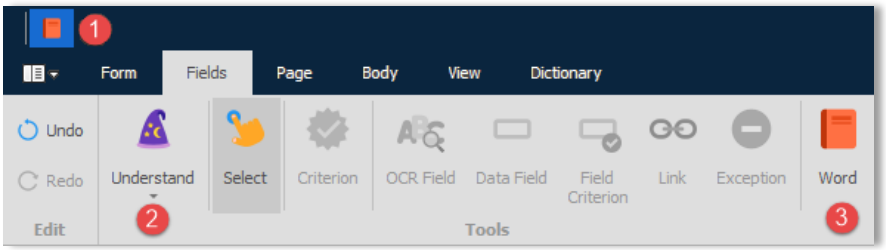
Remove

Select a Criterion, Data Field or Body Field in the Dictionary and click **Remove** (11) to remove all aliases and language specific regular expressions for the selected Term. A warning message will be shown stating before all translations, aliases and language specific regular expressions are lost.

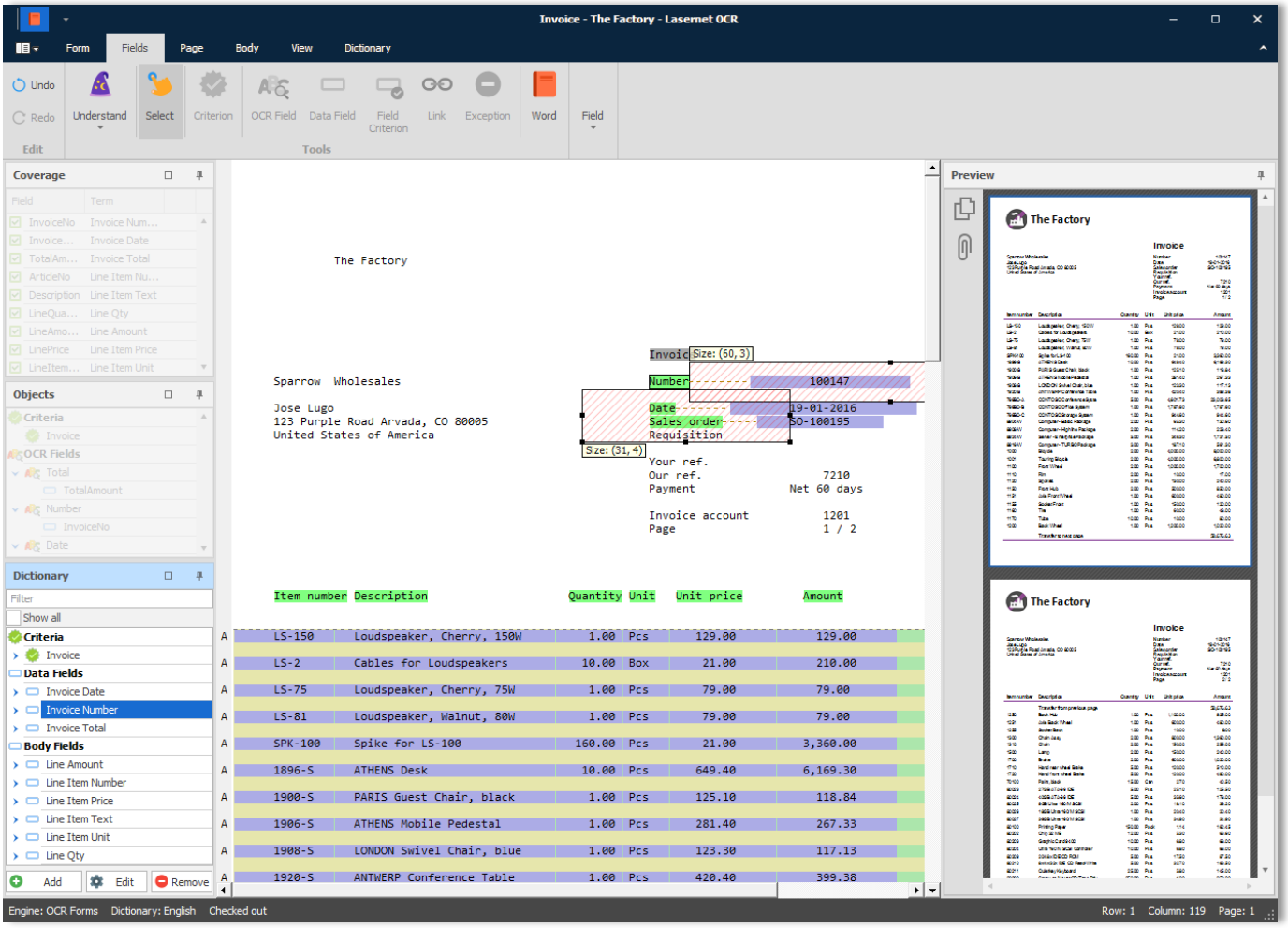


Understand tool

In Dictionary (1) mode there are tools to manage the Dictionary.

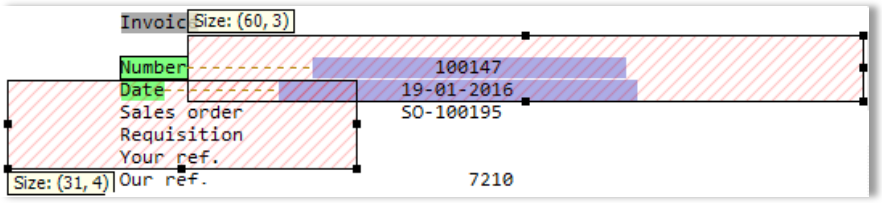


Click the Understand (2) tool and any Alias, connected to a Term that matches a string in the loaded document, will be highlighted in green boxes.

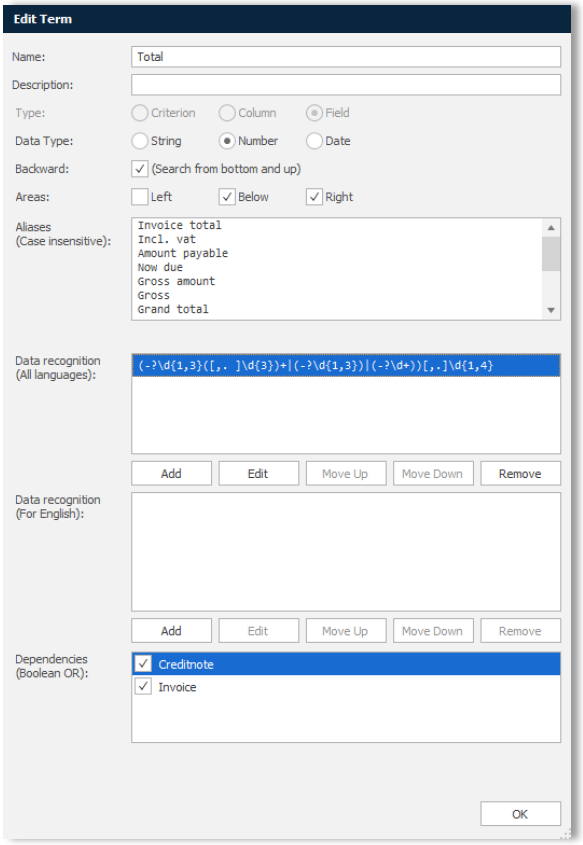


Aliases connected to Terms must be present in the loaded document as well as a Data Recognition (regular expression) in a specified area relative to an Alias. A Data recognition can be set for all languages or a specific language.

Click on any of the found Terms (green boxes) to view the location of area(s) where data recognition is running for the selected object. Drag and drop one of the anchor points to increase or decrease the size of the area for data recognition.



Double-click on the green item to edit the properties of the Term.



Name Name of Term that is stored in the Dictionary.

Description User-defined Description of Term.

Type Criterion defines that the Term is used to create a Criterion object that can recognize an OCR Form.

Column defines that the Term is used to automatically create OCR and Data Fields in the body of the OCR Form.

Field defines that the Term is used to automatically create OCR and Data Fields in the header and footer of the OCR Form.

Data type Selected Data type will automatically be assigned to the OCR Field.

Backward

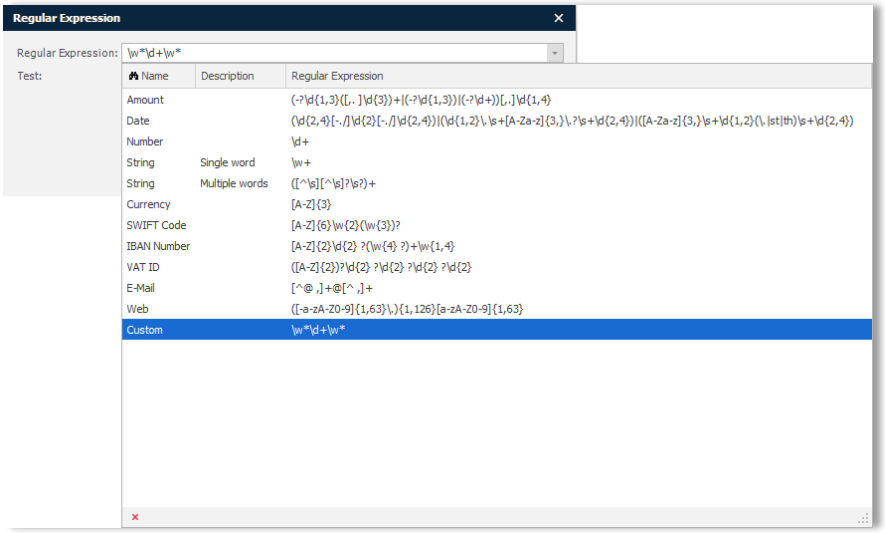
If selected, the search direction is from bottom upwards. We recommend this for fields that are always present at the bottom of a document, for example, the total amount.

Aliases

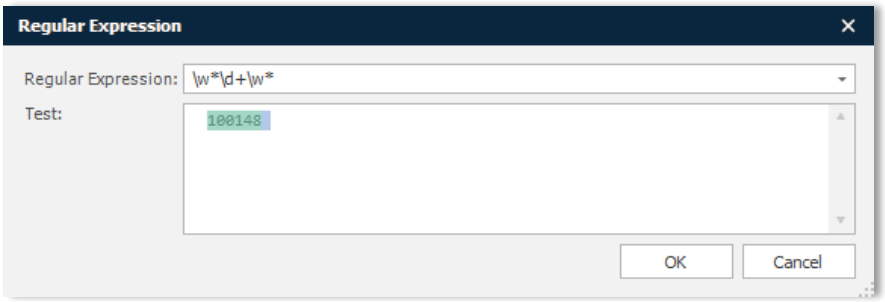
One of the Aliases in the list must appear in an OCR Form to bind a Term to an OCR Field. The text strings are case insensitive and the first found string will bind the Term to an OCR Field if Data recognition is true.

Data recognition

Regular expressions are used for Data recognition either for all languages or a specific language. The field has a list of pre-defined regular expressions to help you set up data recognition for common field types like amounts, dates, numbers, single words, multiple words, currency, SWIFT code, IBAN Number, VAT ID, E-mail and Web. You can customize your own regular expression(s) to handle other types of data recognition.



You can always test various combination of text strings to validate a regular expression. Strings marked with a green background show that the test string has been successfully recognised.



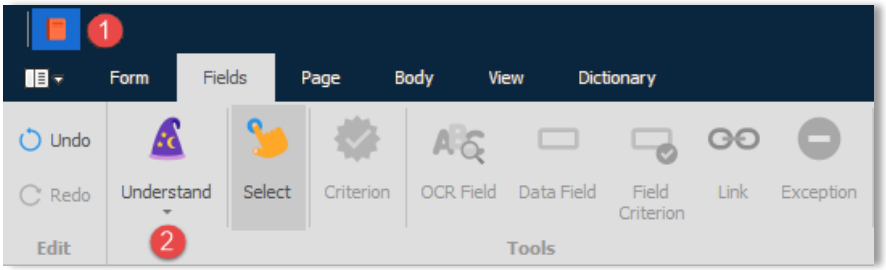
Dependencies

For the most accurate recognition of a Term you can set up Dependencies to specific document type(s). This is often useful if you have an OCR solution running for several Document types and you want to ensure that the Term is only found on a specific type of document.

If you want to add additional words to the Dictionary you can use the Word (3) tool and mark words located in the Document.

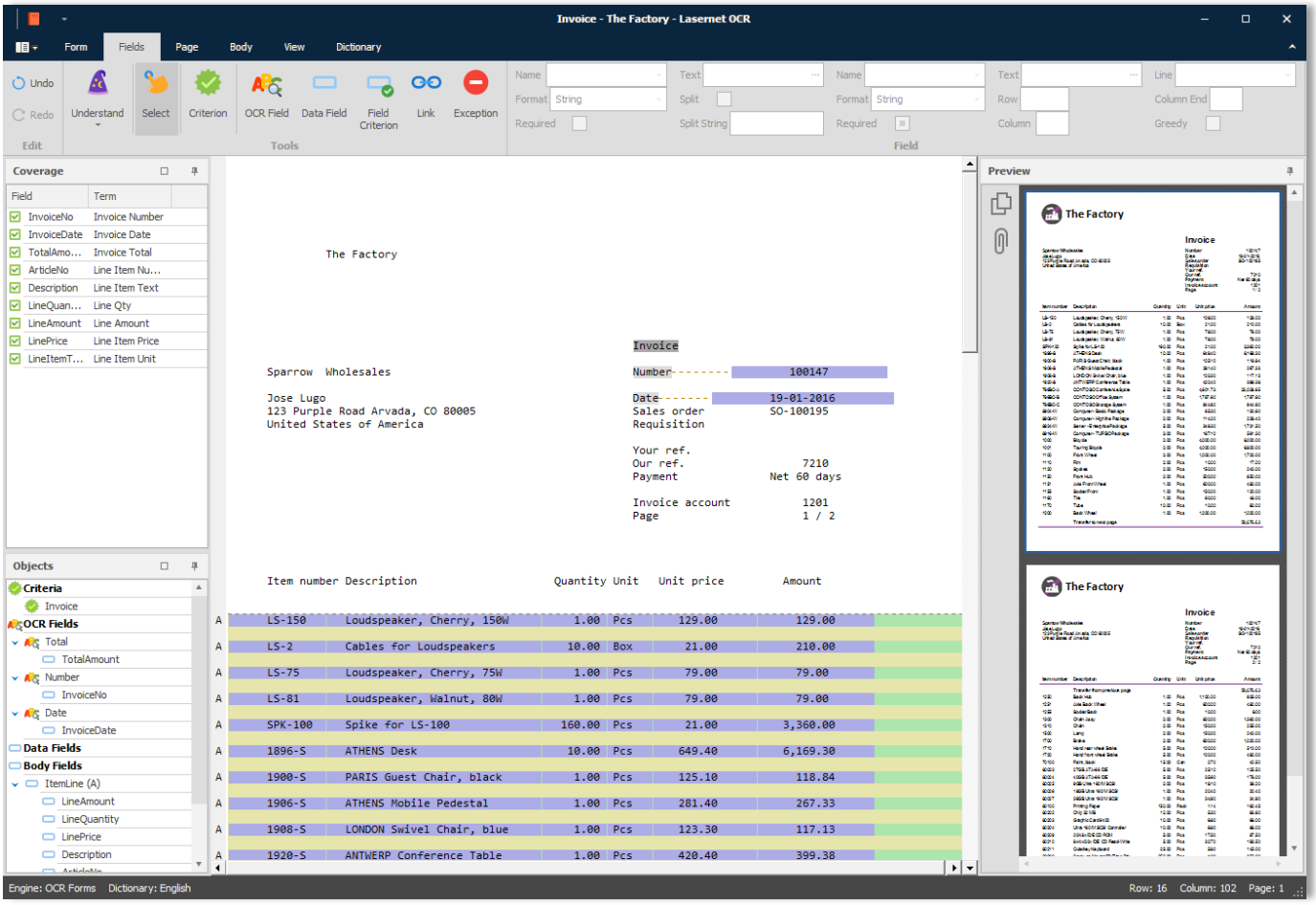
3.8.2 Create OCR Forms with Dictionary

Turn off the Dictionary (1) mode and tools used to manage the Dictionary will be disabled.



You can always switch to Dictionary (1) mode and add new Terms and Aliases to the Dictionary. Existing objects added to the OCR Form will not be lost.

Click the Understand (2) tool and any Alias, connected to a Term that matches a string in the loaded document, will be created as Criteria, OCR Fields and Data Fields.



The Coverage window will show which fields have been automatically covered using the rules in the Dictionary.

The Objects window will show the property values for each object more in detail. If there is data in the document that has not been found, you can add the OCR Fields and Data Fields manually to finish the data extraction.

It is important to ensure that the objects which are automatically found are connected to the expected OCR and Data Fields and have the expected formatting properties. If the data is not extracted as expected, you must either manually correct the OCR Form objects or switch to Dictionary mode to optimize the contents of the Dictionary.

We recommend that you work with a checked-in version of the Dictionary to ensure you are using the latest revision.

3.9 Convert OCR Engine XML to your own workflow format

XML formats created by the OCR Engine are produced to a standard format and are not specific to any workflow system.

It is possible to convert the XML format, created by the OCR Engine, to one that is supported by your workflow system. The configuration and design can be done in the Lasernet Form Editor or Lasernet XML Transformer and processed by the Lasernet Form Engine. Please ensure that you have the required XML Input and Output licenses for Lasernet.