Aquaforest

Searchlight Tagger
Reference Guide

Aquaforest

Making Content Findable

Searchlight Tagger
Reference Guide

Searchlight

Version 1.2

April 2022

Content

1 Product Overview

Aquaforest Searchlight Tagger is a tool that further enhances findability and classification of documents in SharePoint by automatically generating and tagging metadata based on the contents of the documents via rules, taxonomies, barcodes, PDF forms, XMP and integration with NLP services.

1.1 The Business Problem: Drowning in Data, Thirsting for Information

According to an Enterprise Search and Findability survey conducted by Findwise in 2016, more than one-third of respondents stated that it is difficult for users to find information in their organisations and two-thirds of respondents stated that more than 50% of employees are dependent on good findability in their daily work.

With the ever-increasing growth of data being stored to document stores such as Microsoft SharePoint and the increased expectations of good findability, there is a need for a solution to automatically enrich the *(raw)* data by extracting valuable information from them, which can then be used to enhance findability – a critical need for business success.

The extracted information can be added as metadata (also known as tagging) to the documents in SharePoint. Metadata is key to improve findability and retrieve accurate and relevant information in SharePoint. Documents stored in SharePoint may often be lacking key metadata required to enable straightforward metadata searches. As a result, when a query is performed, all documents containing the search term are returned, with no possibility of further refining the search results.

Tagging documents with good metadata improves their ranking in search results by prioritising query matches against the metadata (as compared to matches against the text within the documents), thus providing more relevant results. Moreover, the results can be further refined through faceted navigation. With faceted navigation, multiple filters on various additional metadata can be applied incrementally to drill down to get the correct document/information.

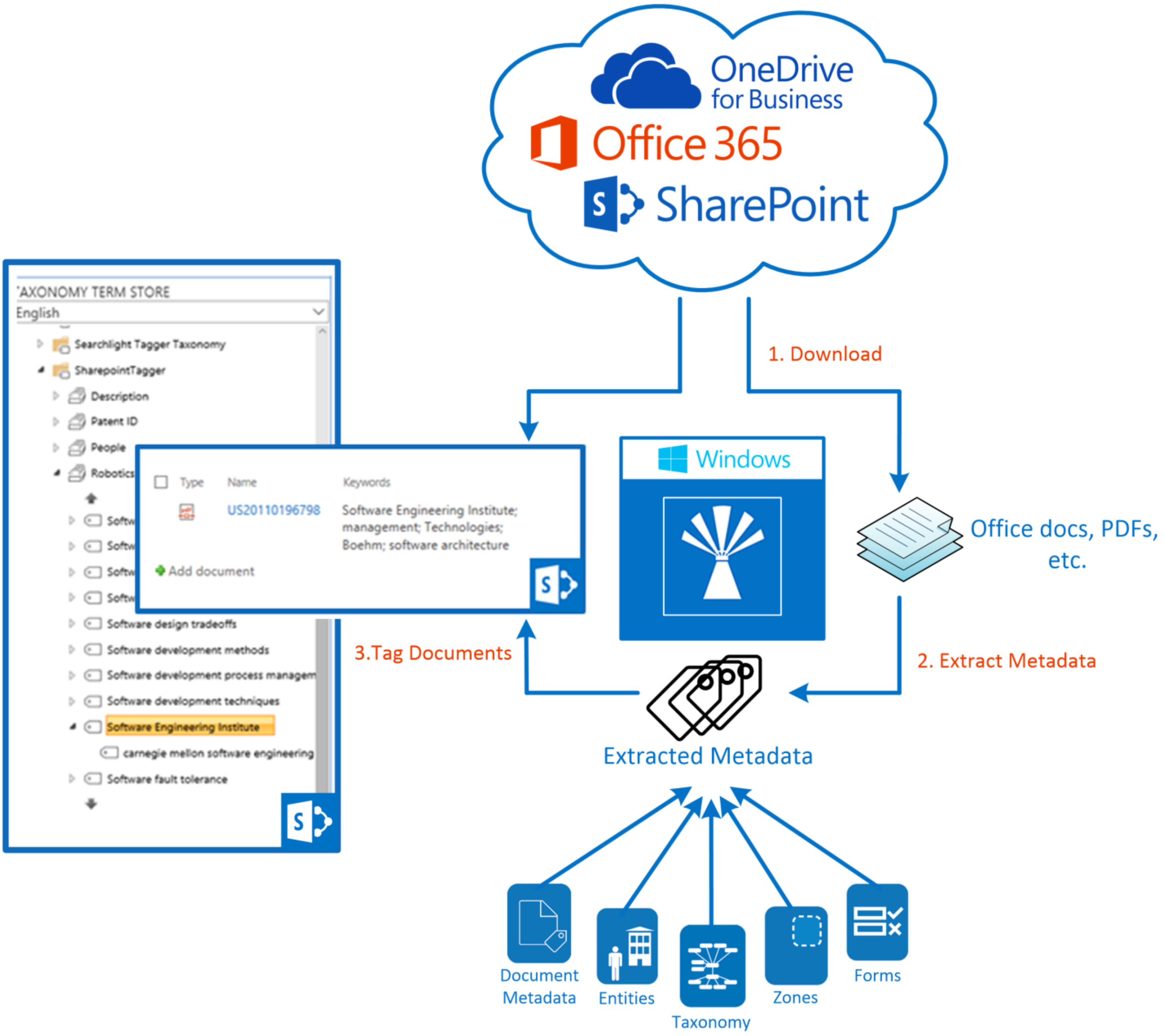
Presently, tagging in organisations is performed manually. According to the SharePoint and Office 365 State of the Market survey by Concept Searching in 2016, 91% of organisations perform some type of manual tagging. However, only 8.4% were satisfied with their tagging accuracy. This is because it is impossible to expect broad sets of employees to accurately tag documents that are often several 100 pages long. Besides, manual tagging is subjective and therefore prone to inconsistencies and ambiguity, not to mention it is also very time consuming. Inconsistent metadata or worst - wrong metadata, negatively affects search results and eventually the business itself.

Consequently, all things considered, automated tagging is the likely practical solution. Automatically generated metadata can be complemented by manual inspections and corrections to improve consistency, accuracy, speed and cost of metadata tagging.

1.2 The Solution: Aquaforest Searchlight Tagger

Aquaforest Searchlight Tagger is a tool that can be configured to automatically extract and/or generate metadata from new and existing documents in SharePoint and tag them accordingly to further enhance findability and classification. It is a stand-alone client application and can be installed on any computer that can connect to the SharePoint server.

1.2.1 Architecture



In a nutshell, Aquaforest Searchlight Tagger works in 3 main steps:

1. Documents are downloaded from SharePoint to the temporary location defined in Tagger

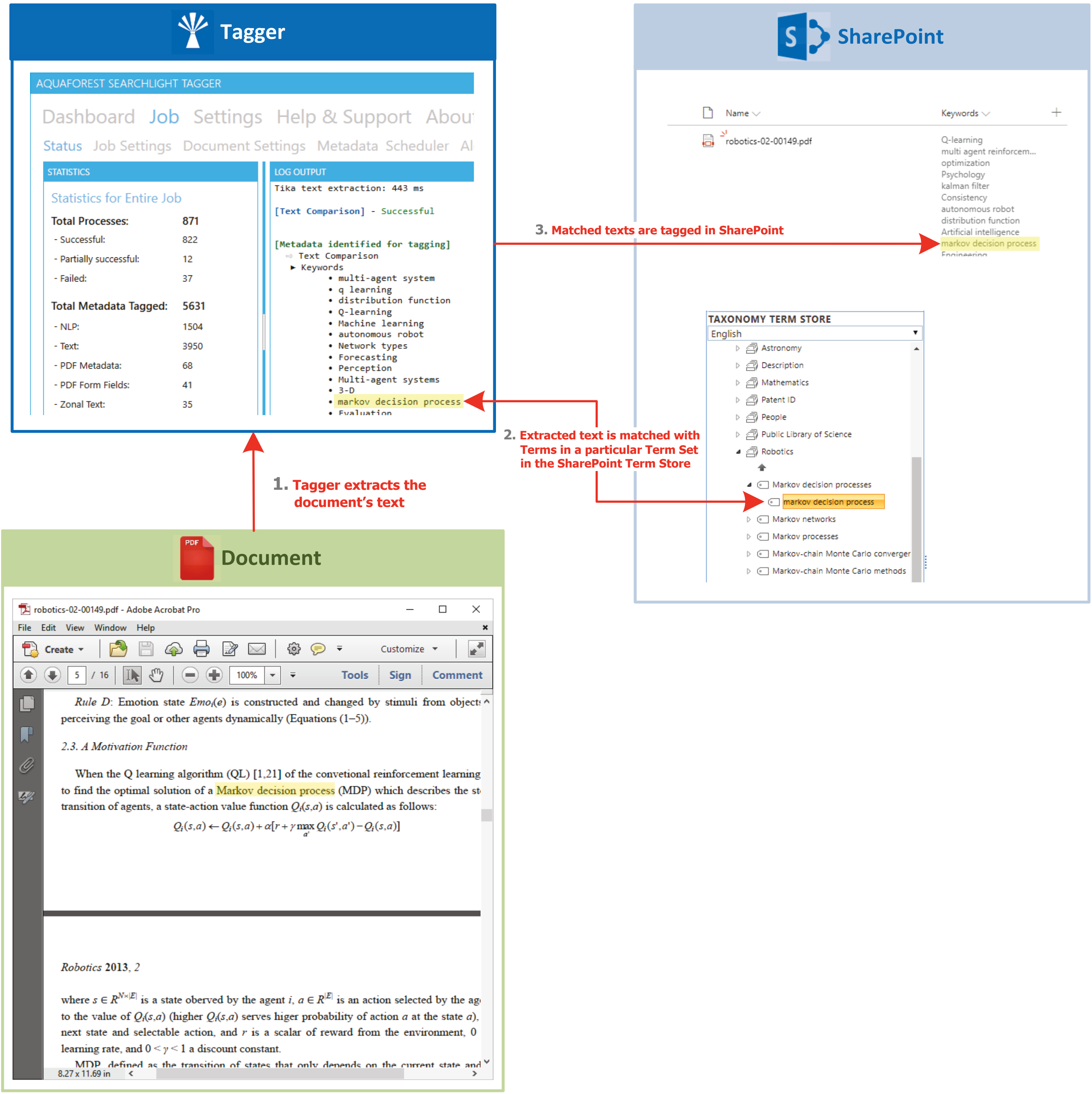
2. Metadata are extracted or generated from the documents based on the extraction type(s) selected and metadata chosen to be extracted. The extraction types are described in the sections below.

3. The documents are then tagged with the extracted metadata from the previous step. If necessary the metadata are added to the Term Store if they are not already present.

   The downloaded documents are deleted after processing.

1.2.2 Taxonomy Matching

Searchlight supports the use of managed metadata and taxonomies for both identifying taxonomy values that should be used to tag the document and is also able to add new taxonomy values if required. Text is extracted from the documents and compared with terms in the Taxonomy Term Store to see if any terms appears in the Text. Only the Terms in the Term Set defined for the selected SharePoint column are compared.



Statistics for Entire Job

Total Processes:	871
- Successful:	822
- Partially successful:	12
- Failed:	37
Total Metadata Tagged:	5631
- NLP:	1504
- Text:	3950
- PDF Metadata:	68
- PDF Form Fields:	41
- Zonal Text:	35

```
Tika text extraction: 443 ms

[Text Comparison] - Successful

[Metadata identified for tagging]
  ⇨ Text Comparison
    ▸ Keywords
        • multi-agent system
        • q learning
        • distribution function
        • Q-learning
        • Machine learning
        • autonomous robot
        • Network types
        • Forecasting
        • Perception
        • Multi-agent systems
        • 3-D
        • markov decision process
        • Evaluation
```

robotics-02-00149.pdf

Keywords: Q-learning, multi agent reinforcem..., optimization, Psychology, kalman filter, Consistency, autonomous robot, distribution function, Artificial intelligence, markov decision process, Engineering

3. Matched texts are tagged in SharePoint

TAXONOMY TERM STORE

English

- Astronomy
- Description
- Mathematics
- Patent ID
- People
- Public Library of Science
- Robotics
  - Markov decision processes
    - markov decision process
  - Markov networks
  - Markov processes
  - Markov-chain Monte Carlo converger
  - Markov-chain Monte Carlo methods

2. Extracted text is matched with Terms in a particular Term Set in the SharePoint Term Store

1. Tagger extracts the document's text

Rule D: Emotion state *Emo_i*(*e*) is constructed and changed by stimuli from object perceiving the goal or other agents dynamically (Equations (1–5)).

2.3. *A Motivation Function*

When the Q learning algorithm (QL) [1,21] of the convetional reinforcement learning to find the optimal solution of a Markov decision process (MDP) which describes the st transition of agents, a state-action value function $Q_i(s,a)$ is calculated as follows:

$$Q_i(s,a) \leftarrow Q_i(s,a) + \alpha[r + \gamma \max_{a'} Q_i(s',a') - Q_i(s,a)]$$

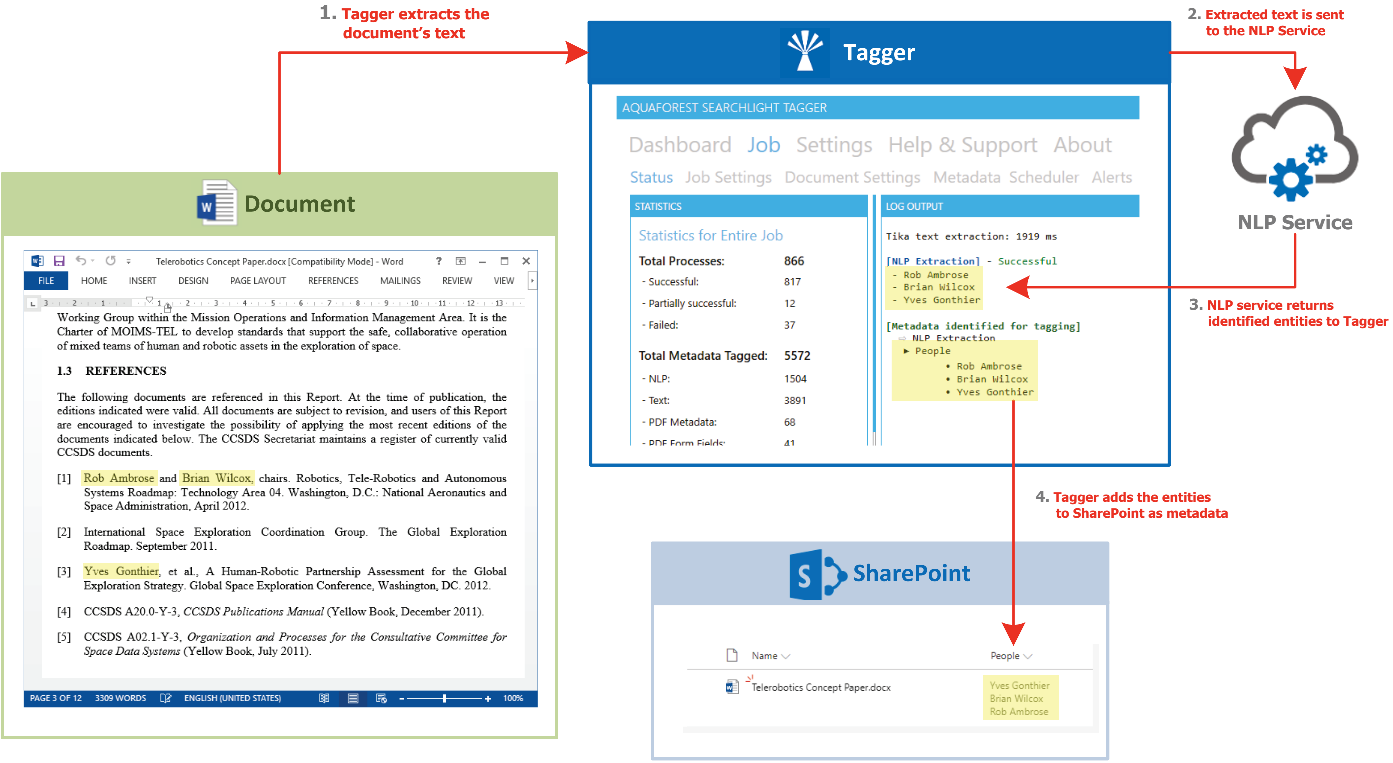
Robotics **2013**, *2*

where $s \in R^{N \times |E|}$ is a state oberved by the agent i, $a \in R^{|E|}$ is an action selected by the ag to the value of $Q_i(s,a)$ (higher $Q_i(s,a)$ serves higer probability of action a at the state a), next state and selectable action, and r is a scalar of reward from the environment, 0 learning rate, and $0 < \gamma < 1$ a discount constant.

MDP defined as the transition of states that only depends on the current state and

1.2.3 Entity Extraction

By integrating with NLP (Natural Language Processing) services, it is able to assign values for Entities such as Location, Person, Company and more. Text is extracted from the documents and passed to the NLP service defined in Tagger. The NLP service will then analyse the text and automatically identify or generate entities to be used as metadata. Entity Extraction is explained in more detail in section 5.1.



1. Tagger extracts the document's text

2. Extracted text is sent to the NLP Service

3. NLP service returns identified entities to Tagger

4. Tagger adds the entities to SharePoint as metadata

Working Group within the Mission Operations and Information Management Area. It is the Charter of MOIMS-TEL to develop standards that support the safe, collaborative operation of mixed teams of human and robotic assets in the exploration of space.

1.3 REFERENCES

The following documents are referenced in this Report. At the time of publication, the editions indicated were valid. All documents are subject to revision, and users of this Report are encouraged to investigate the possibility of applying the most recent editions of the documents indicated below. The CCSDS Secretariat maintains a register of currently valid CCSDS documents.

[1] Rob Ambrose and Brian Wilcox, chairs. Robotics, Tele-Robotics and Autonomous Systems Roadmap: Technology Area 04. Washington, D.C.: National Aeronautics and Space Administration, April 2012.

[2] International Space Exploration Coordination Group. The Global Exploration Roadmap. September 2011.

[3] Yves Gonthier, et al., A Human-Robotic Partnership Assessment for the Global Exploration Strategy. Global Space Exploration Conference, Washington, DC. 2012.

[4] CCSDS A20.0-Y-3, *CCSDS Publications Manual* (Yellow Book, December 2011).

[5] CCSDS A02.1-Y-3, *Organization and Processes for the Consultative Committee for Space Data Systems* (Yellow Book, July 2011).

Statistics for Entire Job

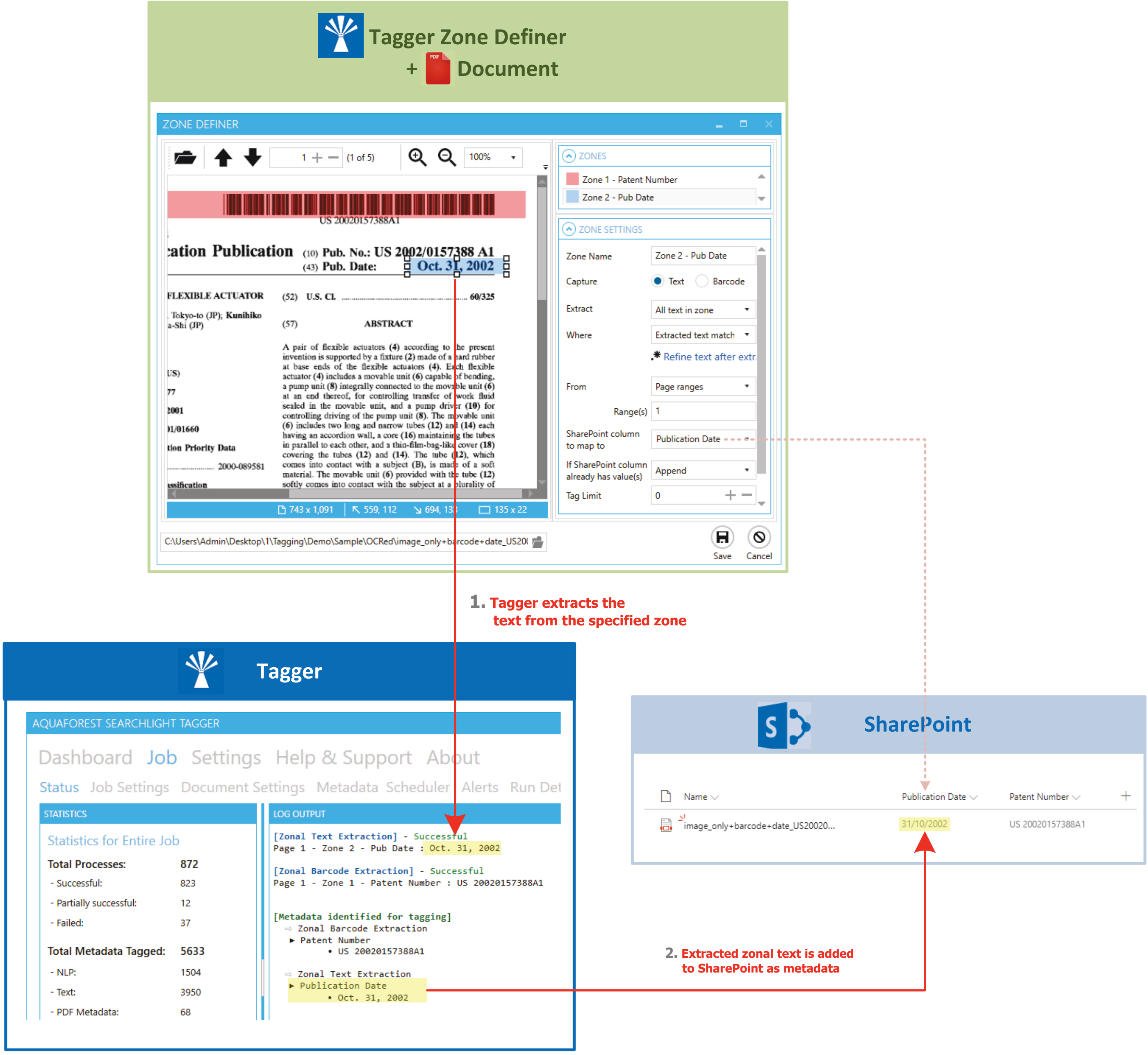
Total Processes:	**866**
- Successful:	817
- Partially successful:	12
- Failed:	37
Total Metadata Tagged:	**5572**
- NLP:	1504
- Text:	3891
- PDF Metadata:	68
- PDF Form Fields:	41

```
Tika text extraction: 1919 ms

[NLP Extraction] - Successful
 - Rob Ambrose
 - Brian Wilcox
 - Yves Gonthier

[Metadata identified for tagging]
  ⇨ NLP Extraction
   ▶ People
        • Rob Ambrose
        • Brian Wilcox
        • Yves Gonthier
```

Name	People
Telerobotics Concept Paper.docx	Yves Gonthier Brian Wilcox Rob Ambrose

1.2.4 Zonal Extraction

It enables zonal extraction of text and barcodes from PDF documents. Over 20 types of barcode can be recognized and the values assigned to Library metadata columns.



Zones:
- Zone 1 - Patent Number
- Zone 2 - Pub Date

Zone Name	Zone 2 - Pub Date
Capture	● Text ○ Barcode
Extract	All text in zone
Where	Extracted text match
	Refine text after extr
From	Page ranges
Range(s)	1
SharePoint column to map to	Publication Date
If SharePoint column already has value(s)	Append
Tag Limit	0

743 x 1,091 | 559, 112 | 694, 13 | 135 x 22

C:\Users\Admin\Desktop\1\Tagging\Demo\Sample\OCRed\image_only+barcode+date_US20(

1. Tagger extracts the text from the specified zone

Statistics for Entire Job

Total Processes:	872
- Successful:	823
- Partially successful:	12
- Failed:	37
Total Metadata Tagged:	5633
- NLP:	1504
- Text:	3950
- PDF Metadata:	68

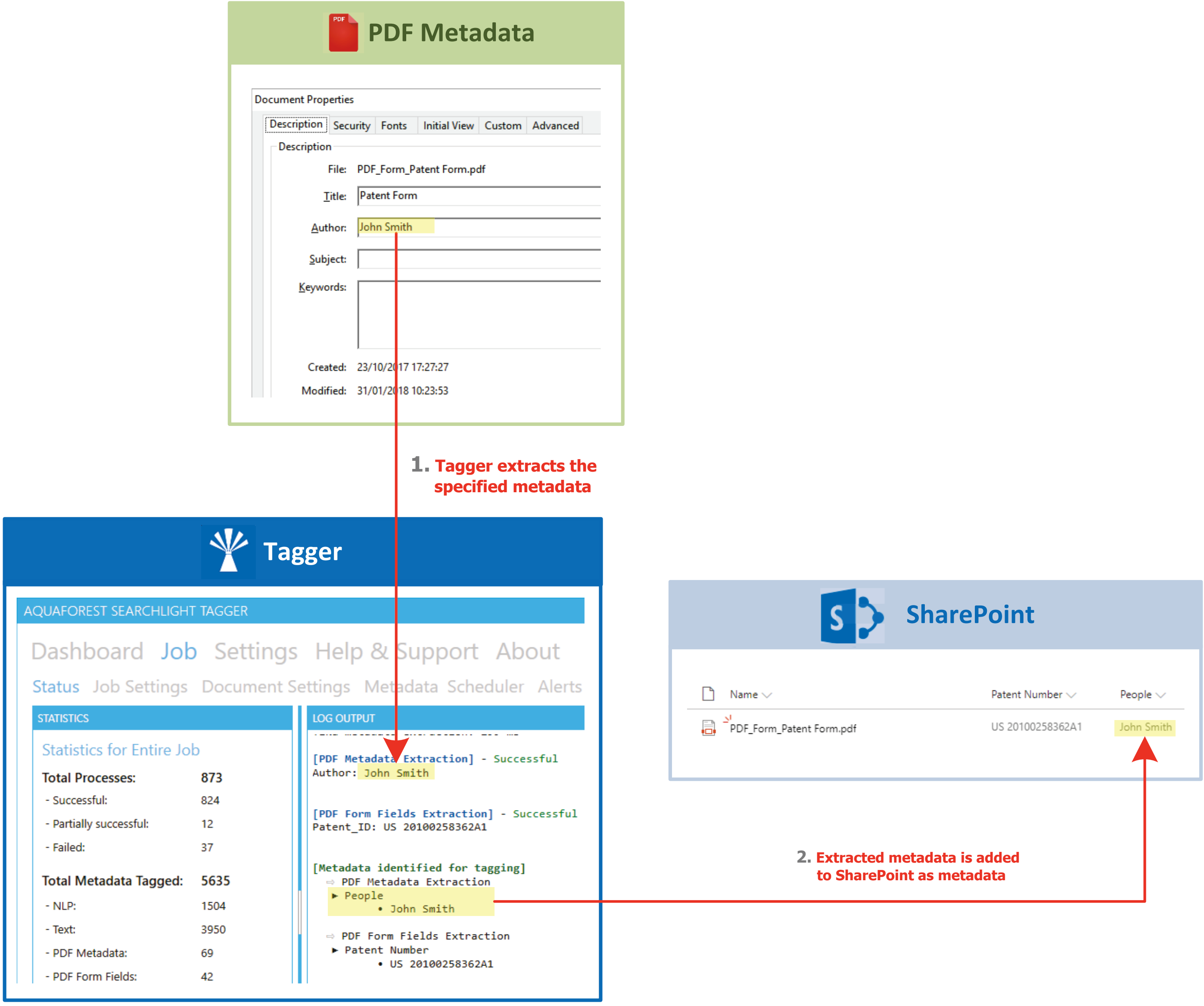
```
[Zonal Text Extraction] - Successful
Page 1 - Zone 2 - Pub Date : Oct. 31, 2002

[Zonal Barcode Extraction] - Successful
Page 1 - Zone 1 - Patent Number : US 20020157388A1

[Metadata identified for tagging]
 ⇨ Zonal Barcode Extraction
  ▸ Patent Number
      • US 20020157388A1

 ⇨ Zonal Text Extraction
  ▸ Publication Date
      • Oct. 31, 2002
```

Name	Publication Date	Patent Number
image_only+barcode+date_US20020...	31/10/2002	US 20020157388A1

2. Extracted zonal text is added to SharePoint as metadata

1.2.5 Document Metadata

Both standard and custom PDF metadata can be extracted and assigned to SharePoint columns. This can also include XMP metadata.



Document Properties

Description

File: PDF_Form_Patent Form.pdf

Title: Patent Form

Author: John Smith

Subject:

Keywords:

Created: 23/10/2017 17:27:27

Modified: 31/01/2018 10:23:53

1. Tagger extracts the specified metadata

STATISTICS

Statistics for Entire Job

Total Processes:	**873**
- Successful:	824
- Partially successful:	12
- Failed:	37
Total Metadata Tagged:	**5635**
- NLP:	1504
- Text:	3950
- PDF Metadata:	69
- PDF Form Fields:	42

LOG OUTPUT

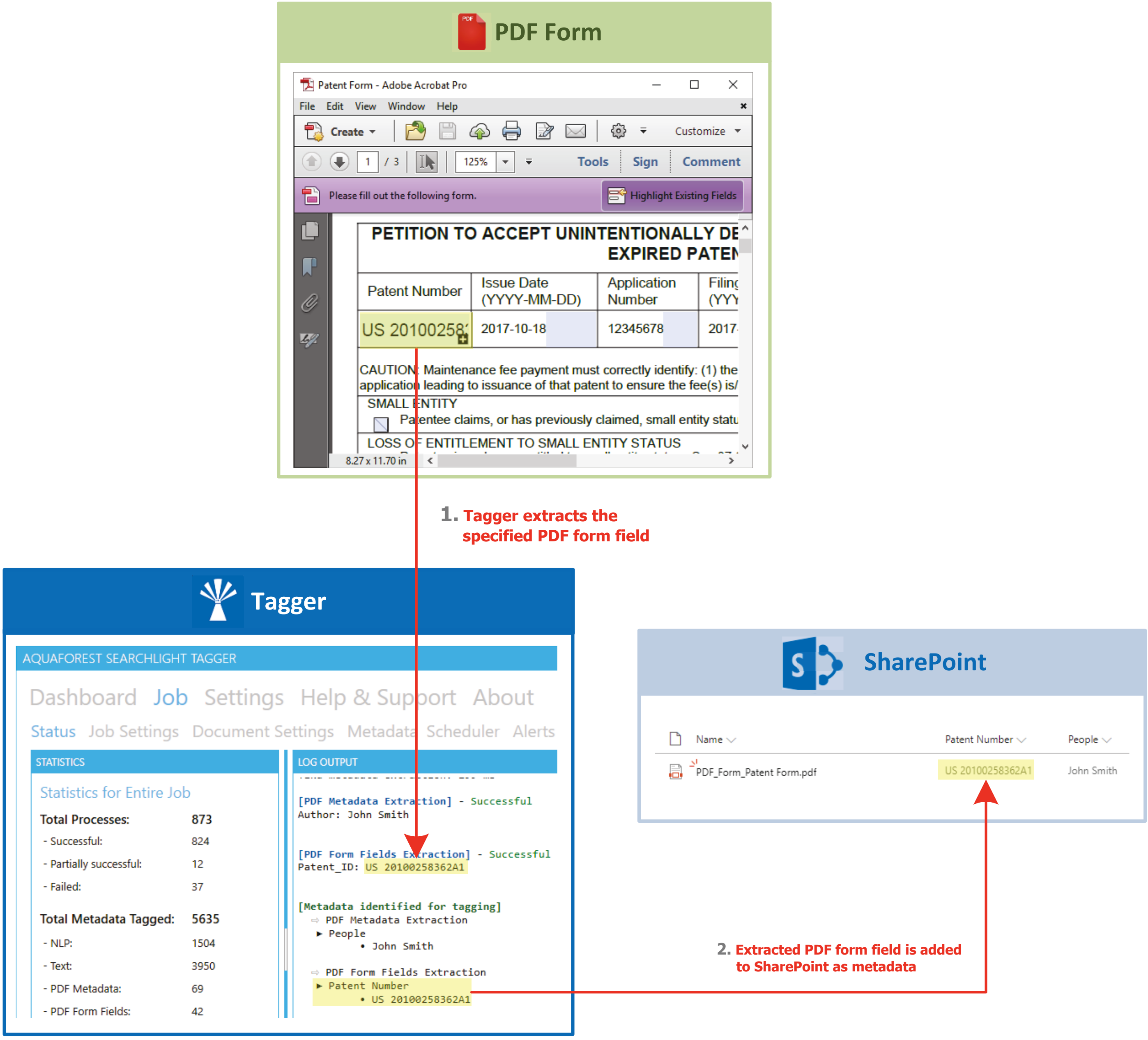
```
[PDF Metadata Extraction] - Successful
Author: John Smith

[PDF Form Fields Extraction] - Successful
Patent_ID: US 20100258362A1

[Metadata identified for tagging]
 ⇨ PDF Metadata Extraction
  ▸ People
        • John Smith

 ⇨ PDF Form Fields Extraction
  ▸ Patent Number
        • US 20100258362A1
```

Name	Patent Number	People
PDF_Form_Patent Form.pdf	US 20100258362A1	John Smith

2. Extracted metadata is added to SharePoint as metadata

1.2.6 PDF Forms

Data from PDF forms can be extracted and each field value assigned to a separate SharePoint column.


PDF Form
Patent Form - Adobe Acrobat Pro
File
Edit
View
Window
Help
Create
Customize
1 / 3
125%
Tools
Sign
Comment
Please fill out the following form.
Highlight Existing Fields
8.27 x 11.70 in

PETITION TO ACCEPT UNINTENTIONALLY DE
EXPIRED PATEN

Patent Number	Issue Date (YYYY-MM-DD)	Application Number	Filing (YYY
US 20100258	2017-10-18	12345678	2017-

CAUTION: Maintenance fee payment must correctly identify: (1) the application leading to issuance of that patent to ensure the fee(s) is/

SMALL ENTITY

□ Patentee claims, or has previously claimed, small entity statu

LOSS OF ENTITLEMENT TO SMALL ENTITY STATUS

1. Tagger extracts the specified PDF form field

Statistics for Entire Job

Total Processes:	873
- Successful:	824
- Partially successful:	12
- Failed:	37
Total Metadata Tagged:	5635
- NLP:	1504
- Text:	3950
- PDF Metadata:	69
- PDF Form Fields:	42

```
[PDF Metadata Extraction] - Successful
Author: John Smith

[PDF Form Fields Extraction] - Successful
Patent_ID: US 20100258362A1

[Metadata identified for tagging]
  ⇨ PDF Metadata Extraction
   ▶ People
        • John Smith

  ⇨ PDF Form Fields Extraction
   ▶ Patent Number
        • US 20100258362A1
```

Name	Patent Number	People
PDF_Form_Patent Form.pdf	US 20100258362A1	John Smith

2. Extracted PDF form field is added to SharePoint as metadata

2 Installation & Licensing

Searchlight Tagger is a standalone client side application (it is not a SharePoint app) that can be installed on any Windows client or server machine that can access your SharePoint instance via a network connection. It does not need to be installed on the SharePoint server.

2.1 System Requirements

Supported Operating Systems	• Windows 7 (x64) • Windows 8 (x64) • Windows 10 (x64) • Windows Server 2008 R2 (x64) • Windows Server 2012 R2 (x64) • Windows Server 2016 • Windows Server 2019
Supported Document Stores	• *SharePoint 2010* • *SharePoint 2013* • *SharePoint 2016* • *SharePoint 2019* • *SharePoint Online (Office 365)*
Disk Space	350 MB
Memory	Minimum 4GB (recommended 8GB)
.NET Framework	4.7.2
Visual C++ Redistributable	Visual C++ 2017 Redistributable (x64)

2.1.1 Licensing

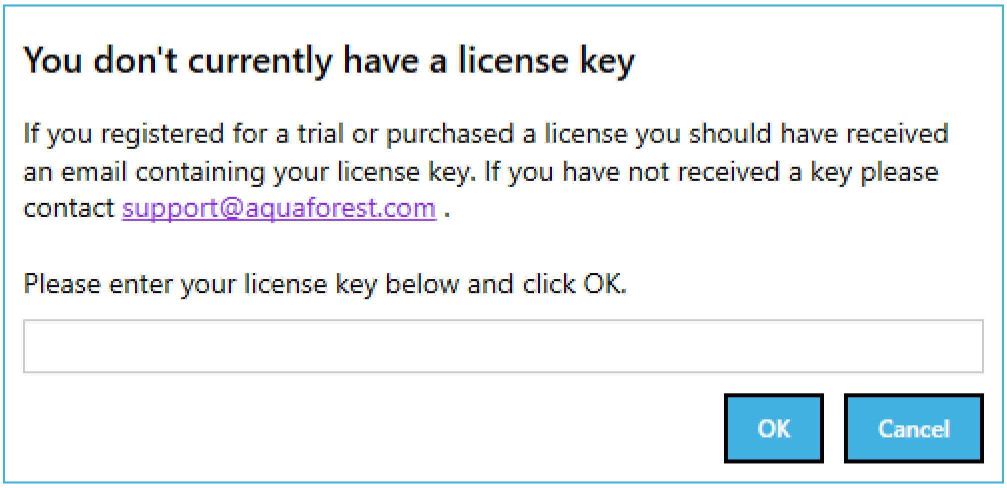
Aquaforest Searchlight Tagger has 3 main licensing levels:

- Single Core
- 4 Cores
- 8 Cores

Trial licenses usually are time limited, that is, it will expire after a particular date or "x" days after installation. They may also limit the number of documents that can be processed.

2.1.2 Entering a license key

Aquaforest Searchlight Tagger will not run without a valid license key. If you do not have a valid license key, you will be prompted to enter one.



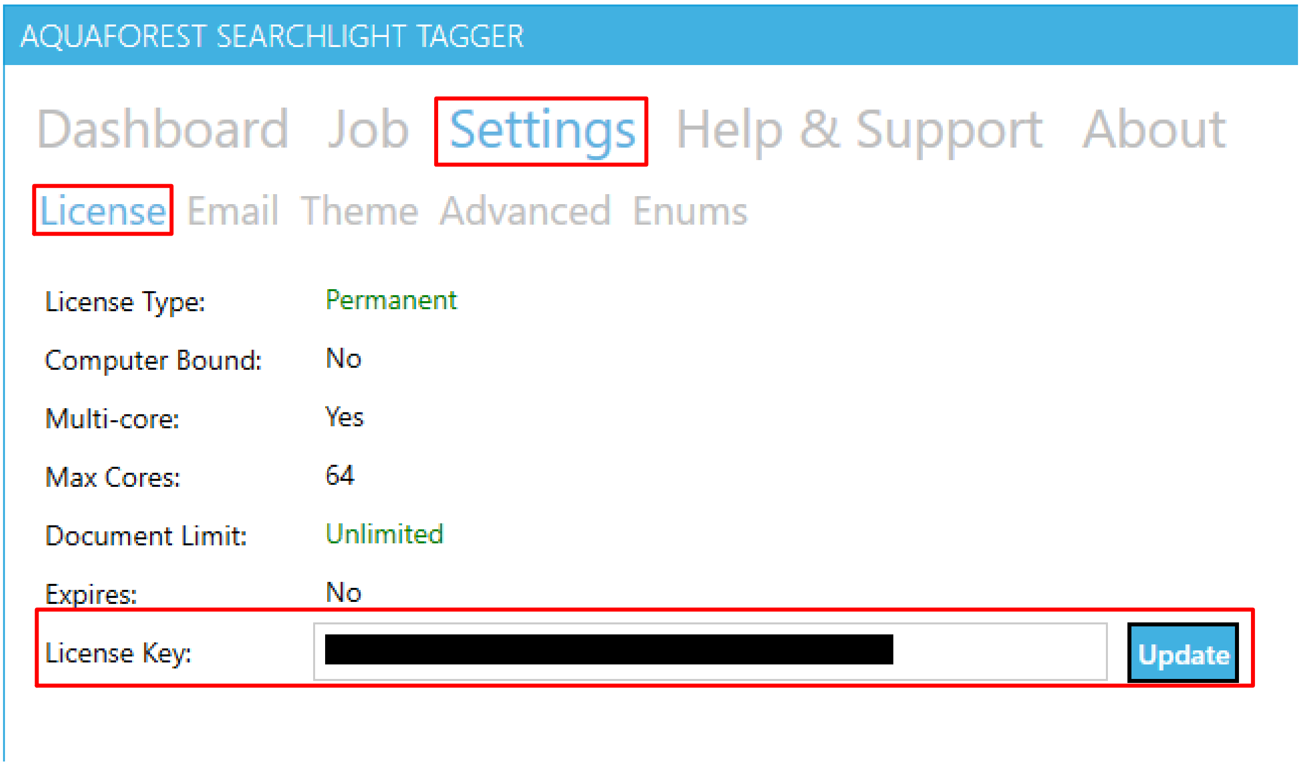
You don't currently have a license key

If you registered for a trial or purchased a license you should have received an email containing your license key. If you have not received a key please contact support@aquaforest.com .

Please enter your license key below and click OK.

Email support@aquaforest.com to request a key if you do not have one.

If you have a valid license key and wish to update it with a new one, go to **Settings** > **License** tab and enter the license in the **License Key** text box and click on **Update**.



License Type:	Permanent
Computer Bound:	No
Multi-core:	Yes
Max Cores:	64
Document Limit:	Unlimited
Expires:	No
License Key:	[redacted]

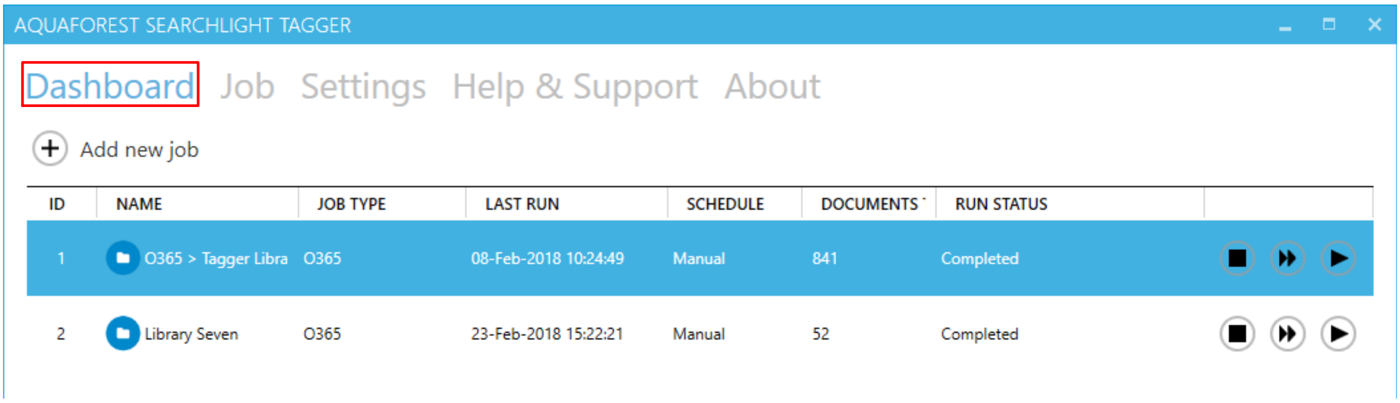
3 Basic Concepts

3.1 Jobs

Aquaforest Searchlight Tagger revolves around the concepts of jobs. A job can be described as an object that has all the settings required to process documents from specific SharePoint locations. It usually consists of the following:

- The location(s) containing the documents that need to be processed.
- Document selection settings to indicate what types of documents to process (docx, pdf, etc.)
- Tagging settings
- Scheduler and alert settings

All jobs are displayed on the **Dashboard** as shown below and the various settings associated with one can be accessed by double-clicking on it.

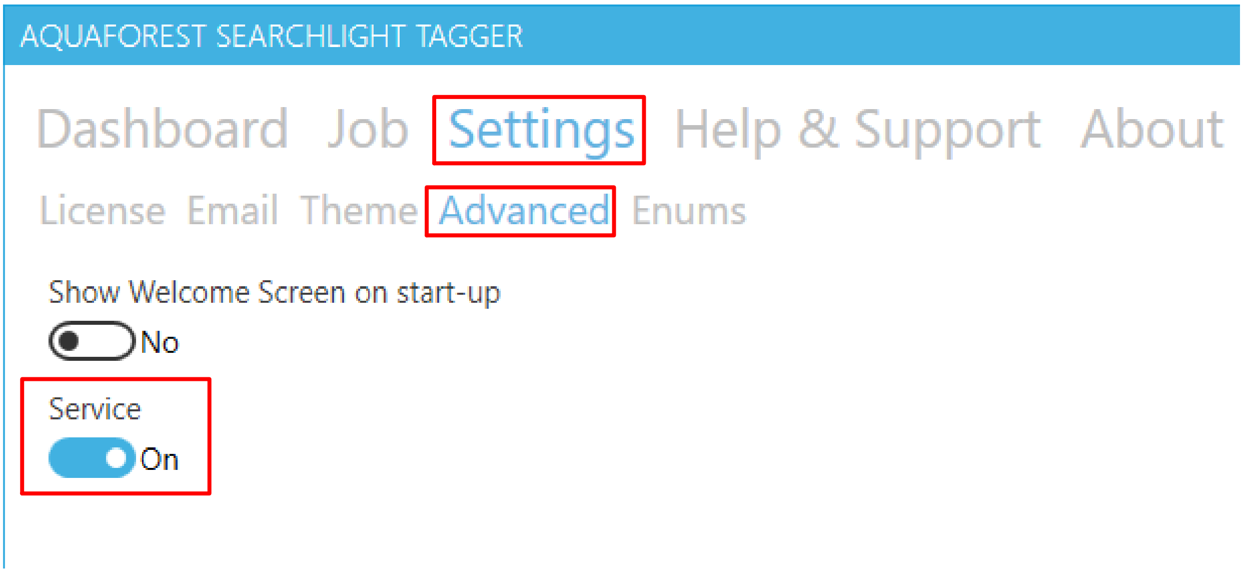


ID	NAME	JOB TYPE	LAST RUN	SCHEDULE	DOCUMENTS	RUN STATUS
1	O365 > Tagger Libra	O365	08-Feb-2018 10:24:49	Manual	841	Completed
2	Library Seven	O365	23-Feb-2018 15:22:21	Manual	52	Completed

3.2 Tagger Service

The Aquaforest Searchlight Tagger service is the heart of the product and controls the execution of all jobs. Without it running, a job cannot be executed. It is also used by the scheduler to automate the processing of jobs at regular time intervals without interfering with other work being performed on the machine it is installed in. It is also used to generate scheduled reports and sending email alerts.

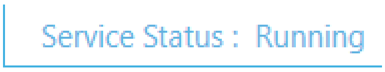
The service can be turned on or off by going to **Settings** > **Advanced** tab.



Show Welcome Screen on start-up
No

Service
On

You can view the current status of the service at the bottom of the Tagger window.



Service Status : Running

3.3 URL format

Below are examples of SharePoint URL formats accepted by Tagger when setting up a job.
NOTE: Make sure the URLs start with "http" or "https"

Site/Web:

- `https://myCompany`
- `https://myCompany/sites/mySite`
- `https://myCompany/sites/mySite/mySubSite`

Document Library:

- `https://myCompany/myLibrary`
- `https://myCompany/sites/mySite/myLibrary`
- `https://myCompany/sites/mySite/mySubSite/myLibrary`

OneDrive for Business

- `https://myCompany-my.sharepoint.com/personal/firstname_lastname_aquaforest_onmicrosoft_com`
- `https://myCompany-my.sharepoint.com/personal/firstname_lastname_aquaforest_onmicrosoft_com/myLibrary`

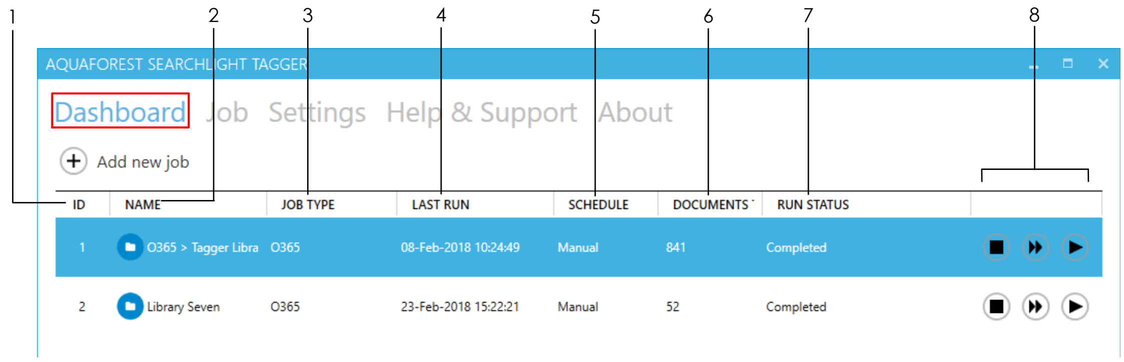
However, even if the full URL is entered (i.e. ending with ".aspx") as shown below, Tagger will try to automatically format it to one of the above accepted formats:

- `https://myCompany/sites/mySite/SitePages/Home.aspx`
- `https://myCompany/sites/mySite/myLibrary/Forms/AllItems.aspx`
- `https://myCompany/sites/mySite/_layouts/15/start.aspx#/myLibrary/Forms/AllItems.aspx`
- `https://myCompany/sites/mySite/Lists/myList/AllItems.aspx`
- `https://myCompany/sites/mySite/_layouts/15/start.aspx#/Lists/myList/AllItems.aspx`
- `https://myCompany-my.sharepoint.com/personal/firstname_lastname_aquaforest_onmicrosoft_com/_layouts/15/onedrive.aspx`
- `https://myCompany-my.sharepoint.com/personal/firstname_lastname_aquaforest_onmicrosoft_com/myLibrary/Forms/AllItems.aspx`

4 Using Tagger

4.1 Dashboard

The dashboard contains all jobs currently defined in Tagger.



ID	NAME	JOB TYPE	LAST RUN	SCHEDULE	DOCUMENTS	RUN STATUS	
1	O365 > Tagger Libra	O365	08-Feb-2018 10:24:49	Manual	841	Completed	■ ⏩ ▶
2	Library Seven	O365	23-Feb-2018 15:22:21	Manual	52	Completed	■ ⏩ ▶

1 The ID of the job (auto generated)

2 The name of the job

3 The SharePoint library type:

- On-premises
- O365

4 The last time the job was run

5 The schedule type of the library:

- Manual
- Daily
- Weekly
- Monthly
- One time

6 The number of documents tagged so far

7 The current status of the job

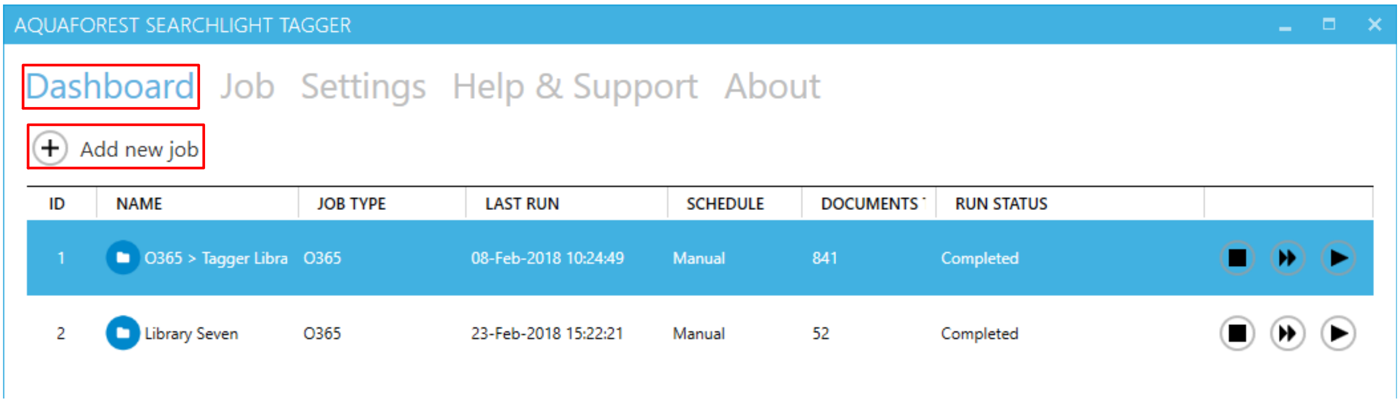
- Not Yet Run
- Completed
- Processing
- Aborted
- Service Error
- License Error
- Database Error
- Document Limit Reached

8

- Abort job ■
- Dry run – perform a test run of the job with the current settings without updating SharePoint ⏩
- Run the job ▶

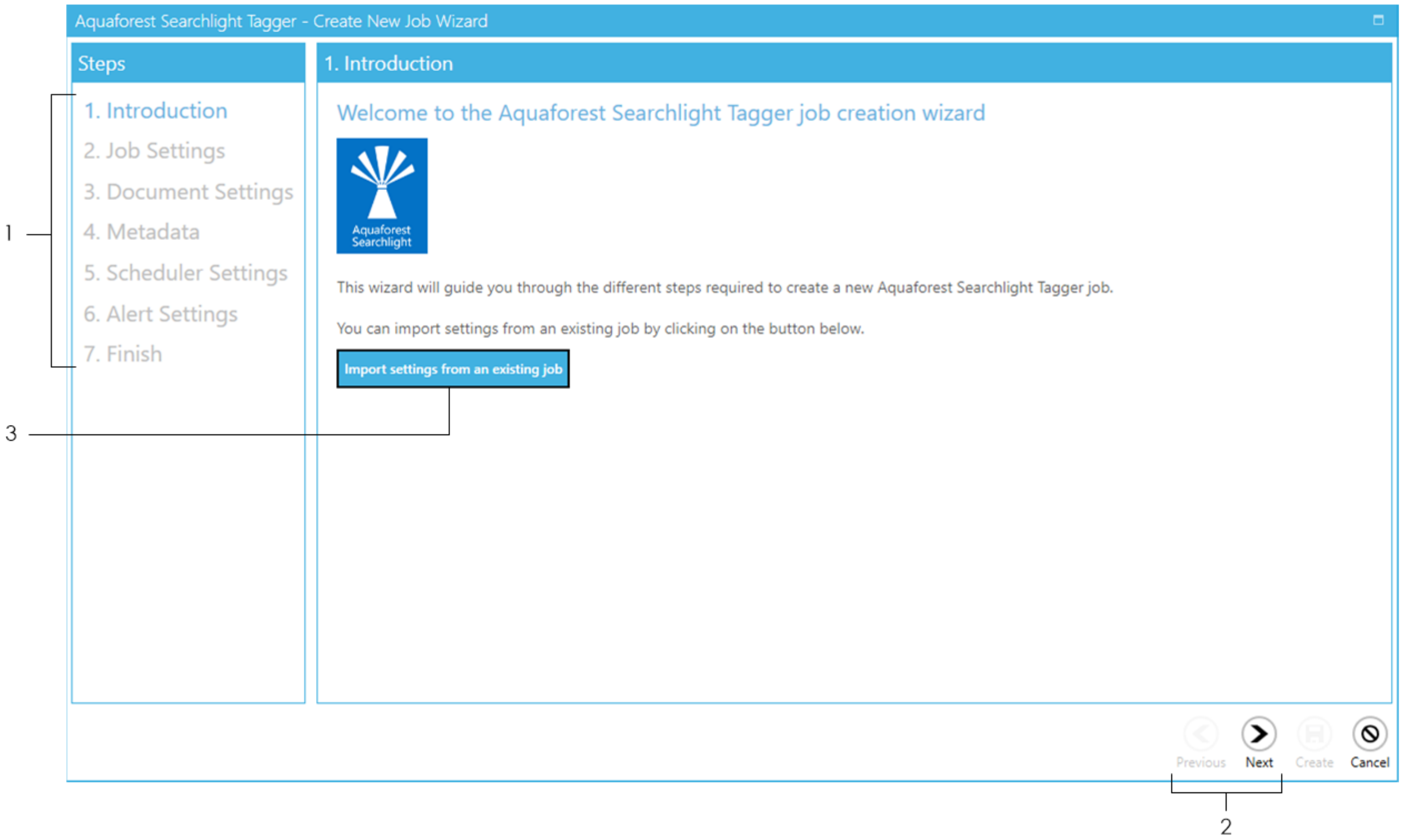
4.2 Creating a job

To create a job go to the **Dashboard** and click on the **Add new job** button.



ID	NAME	JOB TYPE	LAST RUN	SCHEDULE	DOCUMENTS	RUN STATUS
1	O365 > Tagger Libra	O365	08-Feb-2018 10:24:49	Manual	841	Completed
2	Library Seven	O365	23-Feb-2018 15:22:21	Manual	52	Completed

This will launch the job creation wizard, which will guide you through the job creation process step by step.

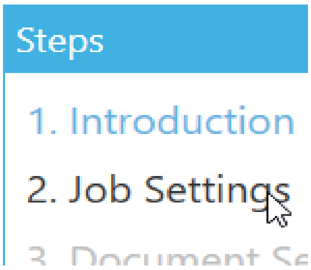


Welcome to the Aquaforest Searchlight Tagger job creation wizard

This wizard will guide you through the different steps required to create a new Aquaforest Searchlight Tagger job.

You can import settings from an existing job by clicking on the button below.

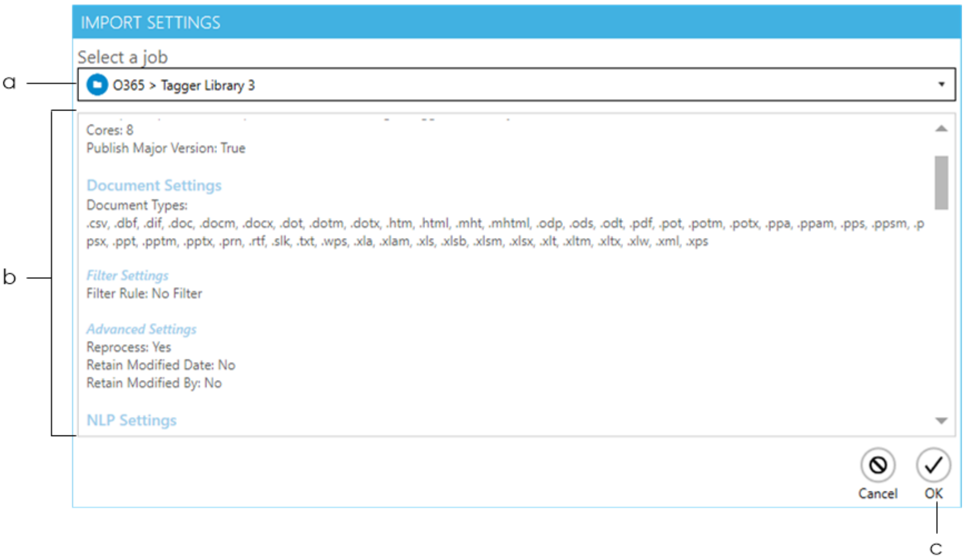
1 Navigate to the various steps by clicking on the step



2 You can also navigate to the other steps by clicking the **Next** or **Previous** buttons at the bottom of the wizard

3 You can choose to **Import settings from an existing job**.

Once the button is clicked, you will be presented with a popup dialog.



Cores: 8
Publish Major Version: True

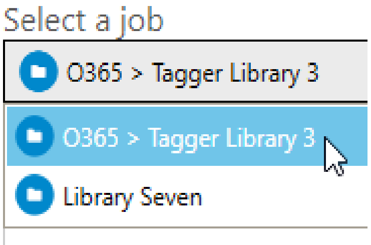
Document Settings
Document Types:
.csv, .dbf, .dif, .doc, .docm, .docx, .dot, .dotm, .dotx, .htm, .html, .mht, .mhtml, .odp, .ods, .odt, .pdf, .pot, .potm, .potx, .ppa, .ppam, .pps, .ppsm, .ppsx, .ppt, .pptm, .pptx, .prn, .rtf, .slk, .txt, .wps, .xla, .xlam, .xls, .xlsb, .xlsm, .xlsx, .xlt, .xltm, .xltx, .xlw, .xml, .xps

Filter Settings
Filter Rule: No Filter

Advanced Settings
Reprocess: Yes
Retain Modified Date: No
Retain Modified By: No

NLP Settings

a. Choose the job you want to copy settings from



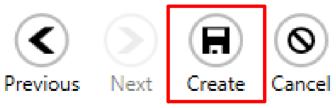
b. This textbox displays a summary of all the settings of the chosen job

c. Click the **OK** button to complete the import

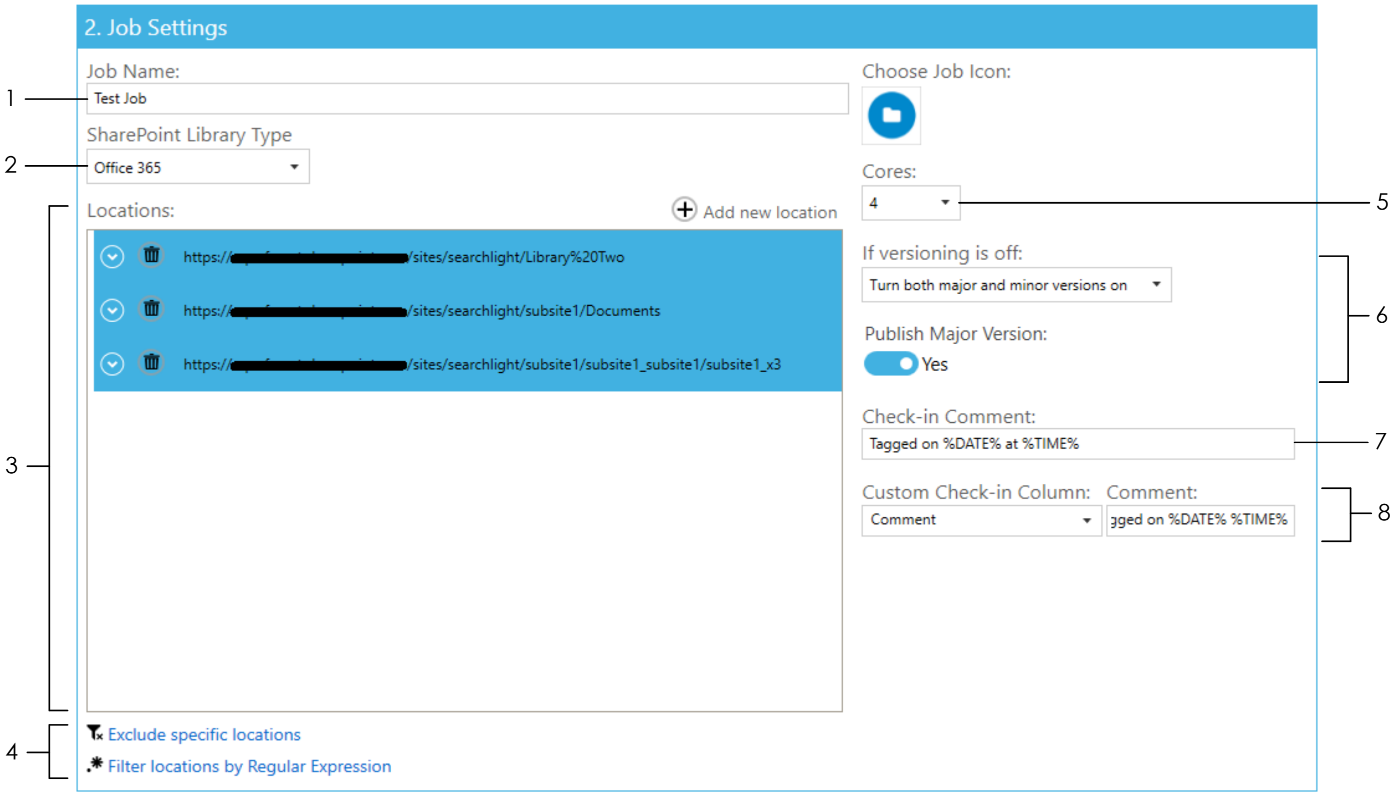
See sections 4.3 to 4.7 for all the settings for all the following steps.

- Job Settings (section 4.3)
- Document Settings (section 4.4)
- Metadata (section 4.5)
- Scheduler Settings (section 4.6)
- Alert Settings (section 4.7)

Once you go through all the steps, you will come to the **Finish** tab, which will show you a summary of all the settings that you have selected. Review them to see if everything are as they should be and finally click on the **Create** button at the bottom of the window.



4.3 Job Settings



2. Job Settings

Job Name: Test Job

SharePoint Library Type: Office 365

Locations:

https://[redacted]/sites/searchlight/Library%20Two
https://[redacted]/sites/searchlight/subsite1/Documents
https://[redacted]/sites/searchlight/subsite1/subsite1_subsite1/subsite1_x3

Choose Job Icon:

Cores: 4

If versioning is off: Turn both major and minor versions on

Publish Major Version: Yes

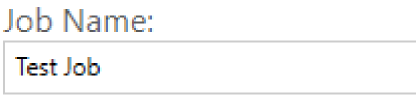
Check-in Comment: Tagged on %DATE% at %TIME%

Custom Check-in Column: Comment

Comment: gged on %DATE% %TIME%

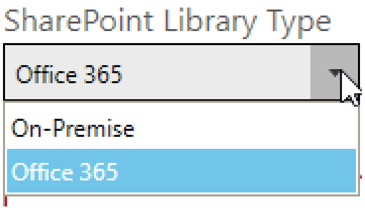
Add new location
Exclude specific locations
Filter locations by Regular Expression
1
2
3
4
5
6
7
8

1 Enter a name for the job



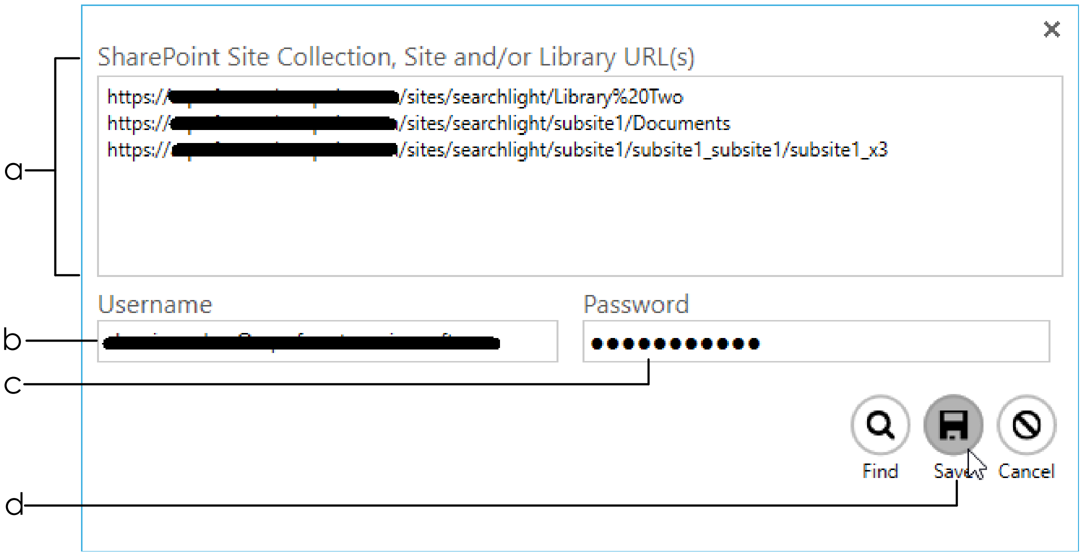
Job Name: Test Job

2 Select the **SharePoint Library Type**



SharePoint Library Type: Office 365

3 Add the SharePoint location(s) by clicking the **Add new location** button. This will open a window as shown below.



SharePoint Site Collection, Site and/or Library URL(s)

https://[redacted]/sites/searchlight/Library%20Two
https://[redacted]/sites/searchlight/subsite1/Documents
https://[redacted]/sites/searchlight/subsite1/subsite1_subsite1/subsite1_x3

Username: [redacted]

Password: ●●●●●●●●●●●

a. Enter the location(s) of the SharePoint site(s) and/or library (ies) you want to process. To enter multiple locations, add each one in a new line.
b. Enter the username to use to access the SharePoint locations. The user should have permissions to modify the locations.
c. Enter a valid password
d. Click on **Save**.

4 You can also filter the locations further by only including or excluding certain locations. This is useful if you are processing a whole site collection and want to excluded specific locations and/or include only specific sites or libraries.

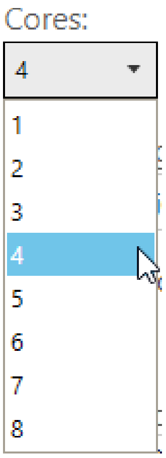
There are 2 ways to filter locations:

a. **Exclude specific locations** – locations that match the specified URL(s) are excluded
b. **Filter locations by Regular Expressions** – locations (URLs) that match the specified regular expressions are included

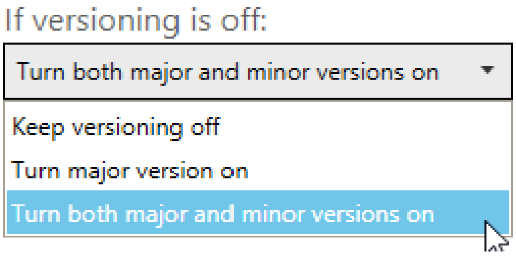
Exclude specific locations
Filter locations by Regular Expression

5 Select the number of cores to use to process documents in parallel. For instance, if 4 cores is specified, Tagger will process 4 documents simultaneously.

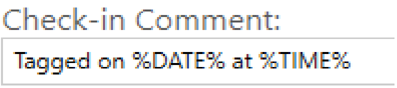
NOTE: The maximum number of cores you can select is limited by your license and the number of processors in the computer where Tagger is installed. To see the amount of cores your license allows, go to **Settings** > **License** tab and check the value for **Max Cores**.


Cores: 4

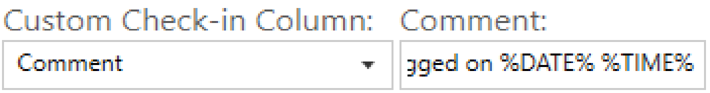
6 Choose whether to turn on versioning if it is turned off on any of the locations.


If versioning is off: Turn both major and minor versions on

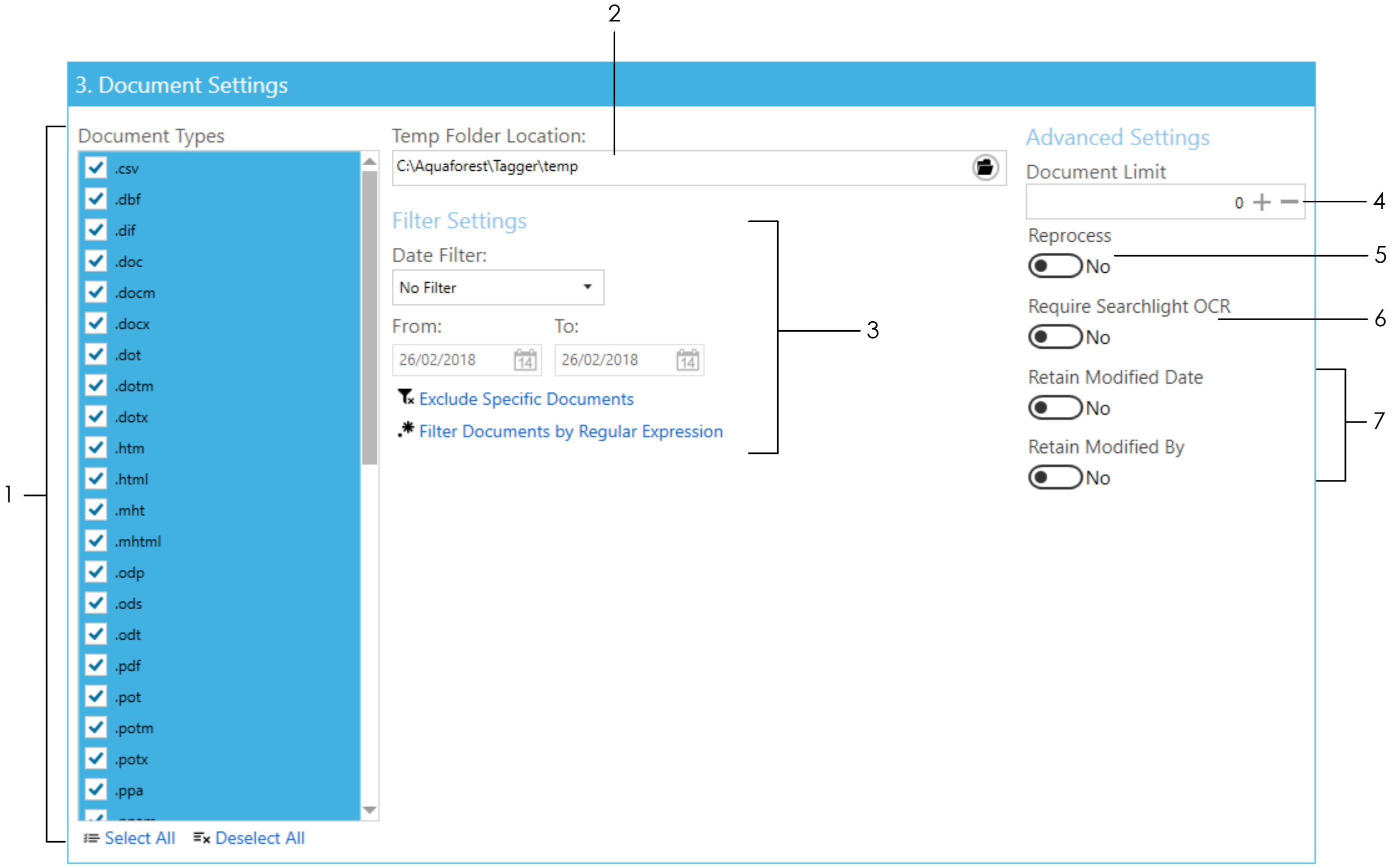
7 You can choose to add a check-in comment to the documents after they are tagged. You can specify the templates ***%DATE%*** and ***%TIME%***, which will be replaced by the date and time the document was tagged.


Check-in Comment: Tagged on %DATE% at %TIME%

8 Optionally, you can also add a custom comment to a custom SharePoint column. The custom SharePoint column must be either of 'Text' or 'Date' type.


Custom Check-in Column: Comment
Comment: ɔged on %DATE% %TIME%

4.4 Document Settings



3. Document Settings

Document Types

- ☑ .csv
- ☑ .dbf
- ☑ .dif
- ☑ .doc
- ☑ .docm
- ☑ .docx
- ☑ .dot
- ☑ .dotm
- ☑ .dotx
- ☑ .htm
- ☑ .html
- ☑ .mht
- ☑ .mhtml
- ☑ .odp
- ☑ .ods
- ☑ .odt
- ☑ .pdf
- ☑ .pot
- ☑ .potm
- ☑ .potx
- ☑ .ppa

Select All Deselect All

Temp Folder Location: C:\Aquaforest\Tagger\temp

Filter Settings

Date Filter: No Filter

From: 26/02/2018 To: 26/02/2018

Exclude Specific Documents

Filter Documents by Regular Expression

Advanced Settings

Document Limit 0

Reprocess No

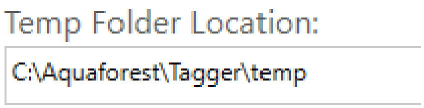
Require Searchlight OCR No

Retain Modified Date No

Retain Modified By No

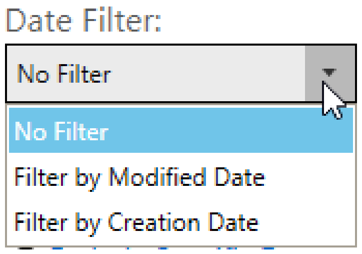
1 Select the document types to process

2 The **Temp Folder Location** is where Tagger temporarily stores downloaded files for processing. Once processing is completed for each document, it is deleted.



Temp Folder Location: C:\Aquaforest\Tagger\temp

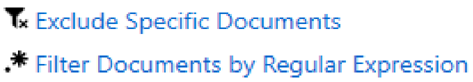
3 There are different options to filter documents:



Date Filter: No Filter

- No Filter
- Filter by Modified Date
- Filter by Creation Date

a. **Date Filter –** Either by modified or creation date. Documents that fall within the specified range are included



Exclude Specific Documents

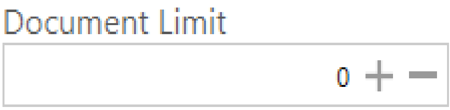
Filter Documents by Regular Expression

b. **Exclude Specific Documents –** documents that match the specified paths are excluded

c. **Filter Documents by Regular Expression** - documents whose properties match the specified regular expressions are included

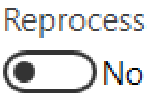
4 You can limit the number of documents to process in each run. This is helpful if you want to process the documents in batches.

Set it to '0' to process all documents.


Document Limit

0

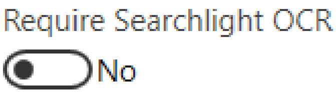
5 Set this to true if you want to re-process documents that have already been tagged. This can be useful if you tagged a document previously using one method (e.g. Zonal) and want to tag it again using another method (e.g. NLP).


Reprocess

No

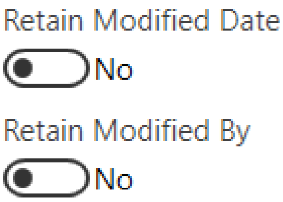
6 This option must be used in conjunction with Searchlight OCR. Set this to true to only process PDF documents that have been processed by Searchlight OCR to make sure they are text searchable before trying to extract metadata.

See section 5.7 for more information about this setting.


Require Searchlight OCR

No

7 You can also **Retain Modified Date** or **Retain Modified By** of the documents in SharePoint so that the Modified Date and Modified By columns will not be changed even after tagging the documents with new metadata.

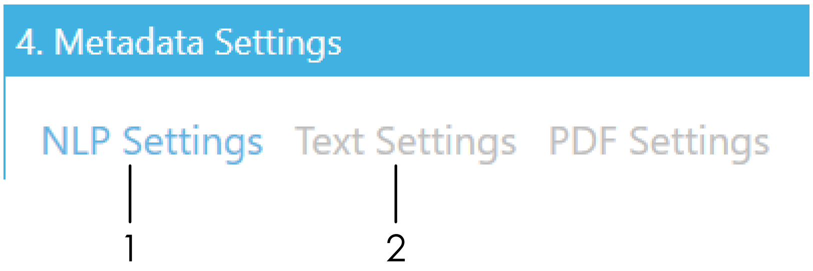

Retain Modified Date

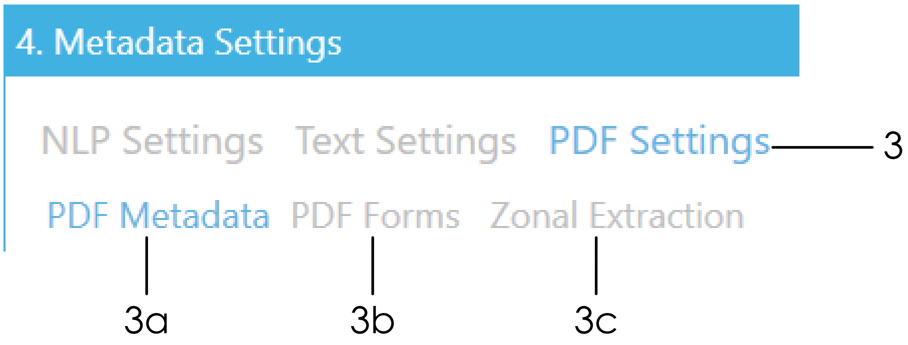
No

Retain Modified By

No

4.5 Metadata





Select how you want to extract metadata from the documents. You can select one or more of the available methods:

1 Entity extraction

2 Taxonomy matching

3 PDF

   a. PDF Metadata

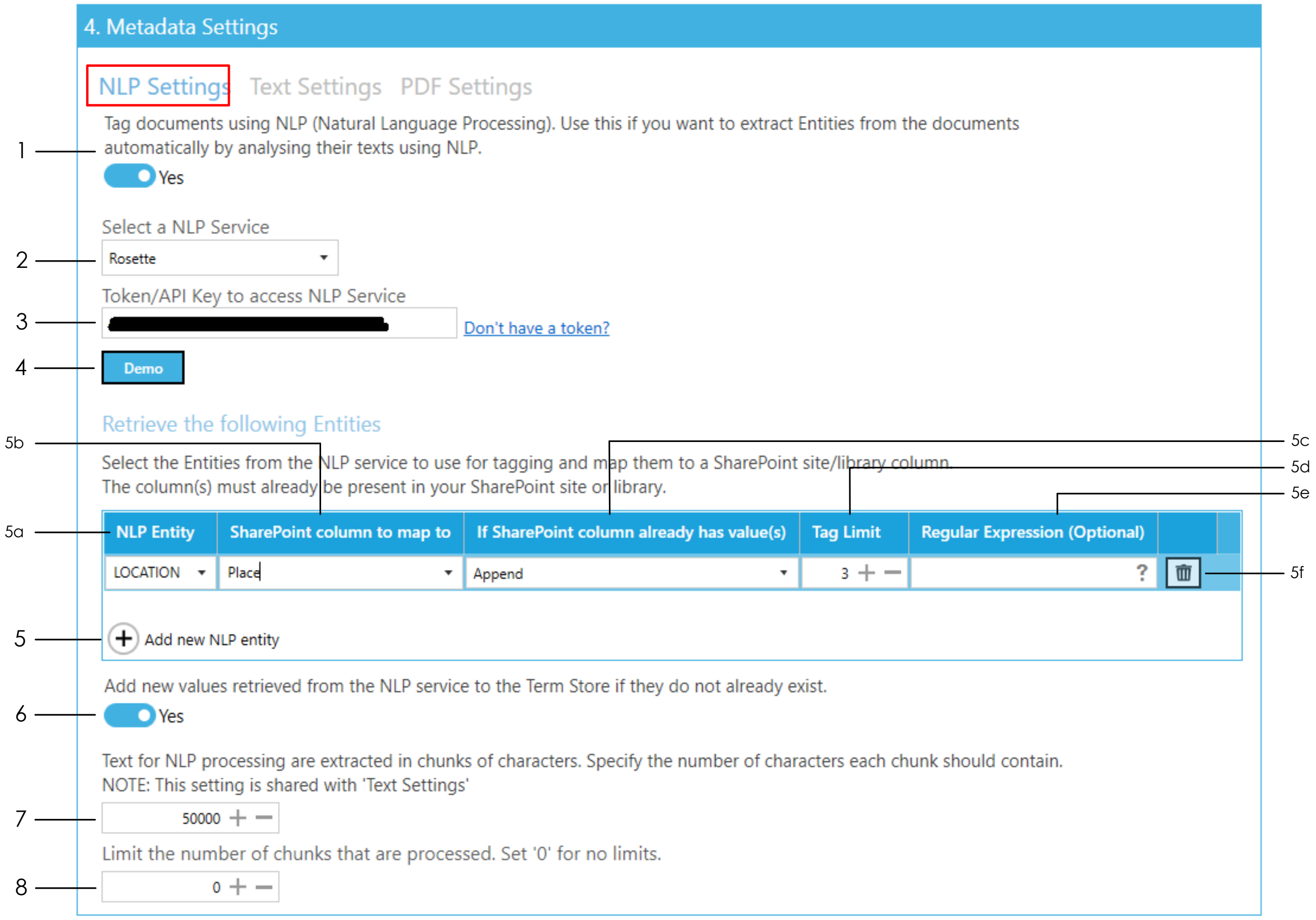
   b. PDF Forms

   c. Zonal

4.5.1 NLP Settings (Entity Extraction)

The following settings deal with automated entity extraction using Natural Language Processing (NLP). It is beneficial to read section 5.1 on Entity Extraction before going through the settings.

These settings are located in **Job** > **Metadata** > **NLP Settings** tab.



Tag documents using NLP (Natural Language Processing). Use this if you want to extract Entities from the documents automatically by analysing their texts using NLP.

Yes

Select a NLP Service

Rosette

Token/API Key to access NLP Service

[redacted] Don't have a token?

Demo

Retrieve the following Entities

Select the Entities from the NLP service to use for tagging and map them to a SharePoint site/library column. The column(s) must already be present in your SharePoint site or library.

NLP Entity	SharePoint column to map to	If SharePoint column already has value(s)	Tag Limit	Regular Expression (Optional)
LOCATION	Place	Append	3	

Add new NLP entity

Add new values retrieved from the NLP service to the Term Store if they do not already exist.

Yes

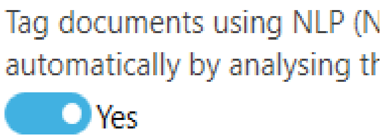
Text for NLP processing are extracted in chunks of characters. Specify the number of characters each chunk should contain.
NOTE: This setting is shared with 'Text Settings'

50000

Limit the number of chunks that are processed. Set '0' for no limits.

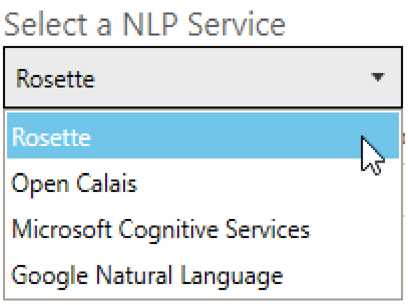
0

1 Enable tagging by NLP



Tag documents using NLP (N
automatically by analysing th
Yes

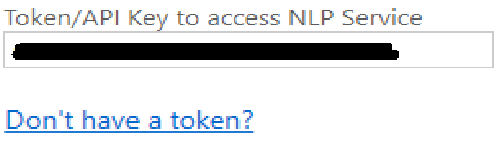
2 Select the NLP Service to use for extracting entities.



Select a NLP Service

Rosette

3 Enter the API key for the selected NLP service.



Token/API Key to access NLP Service

[redacted]

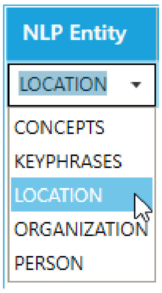
Don't have a token?

If you don't have one, click on the **Don't have a token?** link to sign up to the selected NLP service.

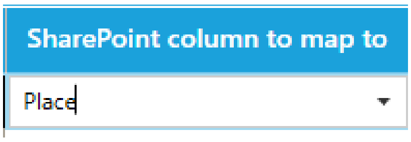
4 You can demo the selected service by clicking the **Demo** button. See [section 5.1.3](#) to see how to use the Demo.

5 To add an NLP entity to extract, click on **Add new NLP entity** button:

a. Select an NLP entity to extract. Each NLP service has its own NLP entities. If you know other NLP entities for a particular NLP service that is not available in the drop-down menu, just type it in.



b. Select or type in the name of the SharePoint column to add the extracted entity to.

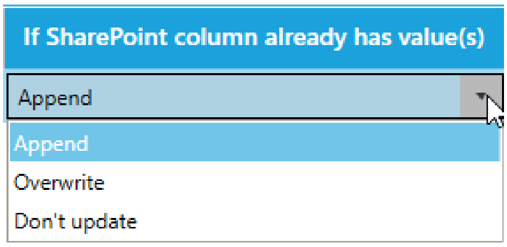


c. Select what to do if the SharePoint column you want to add the extracted entities to already has values:

i. **Append** – the extracted entities will be appended to the existing values of the SharePoint column

ii. **Overwrite** – the existing values of the of the SharePoint column will be deleted and replaced with the new extracted entities

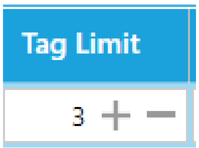
iii. **Don't update** – if the SharePoint column already has values, entity extraction for this column will be skipped



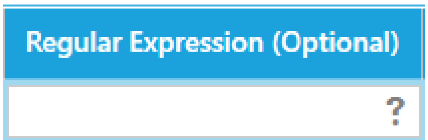
d. Enter a **Tag Limit**. This restricts the number of entities that will be added (tagged) to this SharePoint column. For instance, with the settings displayed in the image above, if 10 'LOCATION' entities are extracted from a document by the selected NLP service, only 3 of them will be added to the SharePoint column.

Set '0' for no limits.

See [section 5.5](#) for more information on tag limits.



e. Optionally, you can specify a [regular expression](#) to ensure that the entities returned by the NLP service are the ones you are looking for. For example, you can specify that the Location starts with a particular word or set of words. To find out more about regular expressions click on the ? icon.

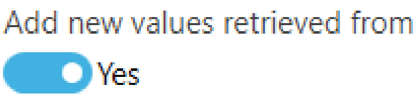


f. Click on the **Delete** button to delete any unwanted rows or to start over.

Repeat the above steps to add more entities.

If any of the SharePoint columns added in step 5 is a Managed Metadata column, it will be associated with an existing Term Set in the Term Store.

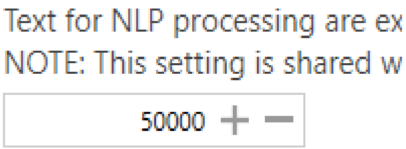
Set this option to 'Yes' if you want to add entities extracted from the NLP service that are not currently present in the Term Set to the Term Store. Otherwise, the entities cannot be tagged and will be skipped.



Add new values retrieved from

Yes

7 In order to send a document to an NLP service, its contents (text) has to be chunked because all NLP services have a limit on the number of characters it can process at any one time, especially if you are using the free option. Normally, the current setting of 50,000 characters should be suitable but if you switch to the premium service of a NLP service, you may be able to increase this value. See section 5.1 for more information.

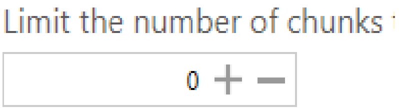


Text for NLP processing are ex
NOTE: This setting is shared w

50000 + −

8 Set this to only process the first 'x' number of chunks. This can be useful if you are processing very large documents and the entities can be extracted on the first few pages.

Set '0' to send the whole document to the NLP service.

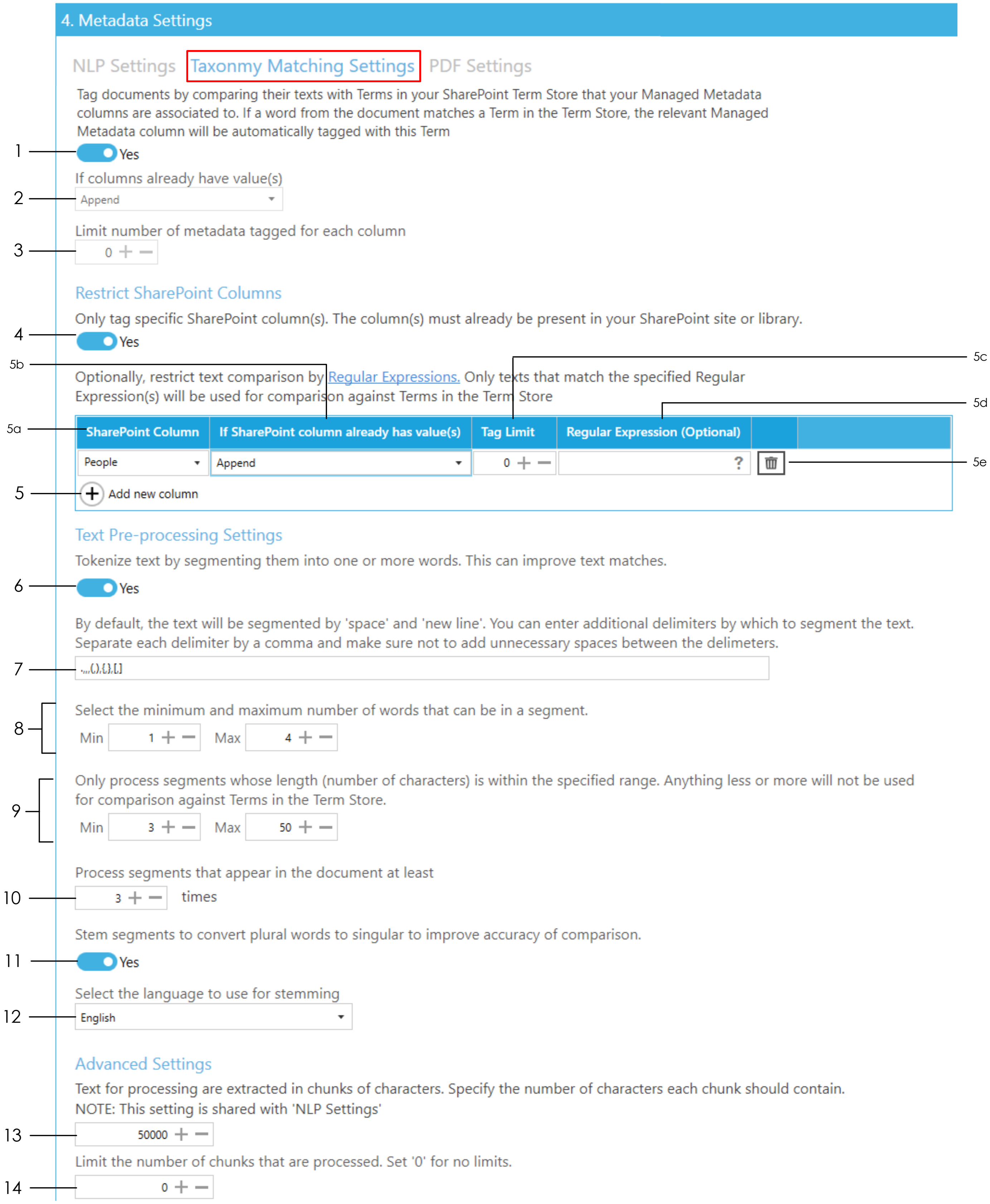


Limit the number of chunks

0 + −

4.5.2 Taxonomy Matching Settings

The following settings is used for taxonomy matching. These settings are located in **Job** > **Metadata** > **Taxonomy Matching Settings** tab.



4. Metadata Settings

NLP Settings | Taxonmy Matching Settings | PDF Settings

Tag documents by comparing their texts with Terms in your SharePoint Term Store that your Managed Metadata columns are associated to. If a word from the document matches a Term in the Term Store, the relevant Managed Metadata column will be automatically tagged with this Term

Yes

If columns already have value(s)

Append

Limit number of metadata tagged for each column

0

Restrict SharePoint Columns

Only tag specific SharePoint column(s). The column(s) must already be present in your SharePoint site or library.

Yes

Optionally, restrict text comparison by Regular Expressions. Only texts that match the specified Regular Expression(s) will be used for comparison against Terms in the Term Store

SharePoint Column	If SharePoint column already has value(s)	Tag Limit	Regular Expression (Optional)
People	Append	0	?

Add new column

Text Pre-processing Settings

Tokenize text by segmenting them into one or more words. This can improve text matches.

Yes

By default, the text will be segmented by 'space' and 'new line'. You can enter additional delimiters by which to segment the text. Separate each delimiter by a comma and make sure not to add unnecessary spaces between the delimeters.

',.,(),{},[]

Select the minimum and maximum number of words that can be in a segment.

Min 1 Max 4

Only process segments whose length (number of characters) is within the specified range. Anything less or more will not be used for comparison against Terms in the Term Store.

Min 3 Max 50

Process segments that appear in the document at least

3 times

Stem segments to convert plural words to singular to improve accuracy of comparison.

Yes

Select the language to use for stemming

English

Advanced Settings

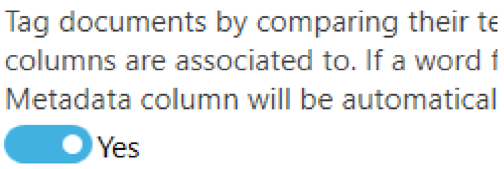
Text for processing are extracted in chunks of characters. Specify the number of characters each chunk should contain.
NOTE: This setting is shared with 'NLP Settings'

50000

Limit the number of chunks that are processed. Set '0' for no limits.

0

1 Enable tagging by comparing Terms in the Term Store

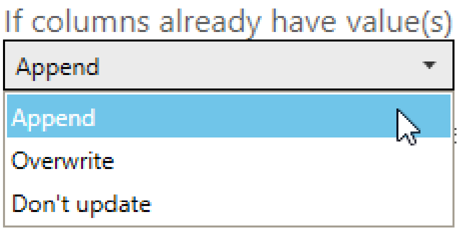


Tag documents by comparing their te
columns are associated to. If a word f
Metadata column will be automatical

Yes

2 Select what to do if the managed metadata SharePoint column already has value(s):

a. **Append** – the identified terms will be appended to the existing values of the SharePoint column
b. **Overwrite** – the existing values of the of the SharePoint column will be deleted and replaced with the new identified terms
c. **Don't update** – if the SharePoint column already has values, text comparison for the column will be skipped



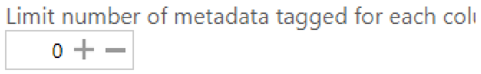
If columns already have value(s)

Append

3 Enter a **Tag Limit**. This restricts the number of term matches that will be added (tagged) to this SharePoint column. For instance, if this is set to 2 and 5 terms are identified from the Term Store, only 2 values will be added to the SharePoint column.

Set '0' for no limits.

See section 5.5 for more information on tag limits.

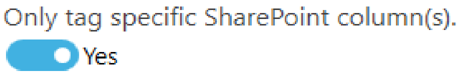


Limit number of metadata tagged for each col

0 + −

4 Set this to 'Yes' to manually specify which column's Term Set to use for comparison. Once this is set to 'Yes', steps 2 and 3 above are overwritten by step 5 below.

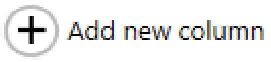
If this is set to 'No', Tagger will identify all the Managed Metadata columns in the specified locations and for each column, it will identify the Term Set associated with it and use those Term Sets for comparison.



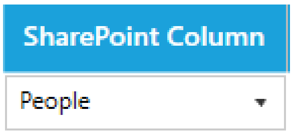
Only tag specific SharePoint column(s).

Yes

5 To add a Managed Metadata column, click on **Add new column** button:



a. Select or type in the name of the SharePoint Managed Metadata column to add the text match to.

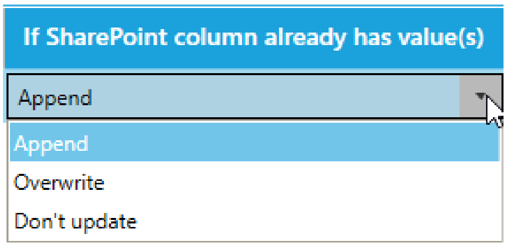


SharePoint Column

People

b. Select what to do if the SharePoint column you want to add the text matches to already has values:

i. **Append** – the identified terms will be appended to the existing values of the SharePoint column
ii. **Overwrite** – the existing values of the of the SharePoint column will be deleted and replaced with the new identified terms



If SharePoint column already has value(s)

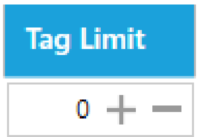
Append

iii. **Don't update** – if the SharePoint column already has values, comparison for the column will be skipped

c. Enter a **Tag Limit**. This defines the number of term matches that will be added (tagged) to this SharePoint column.

Set '0' for no limits.

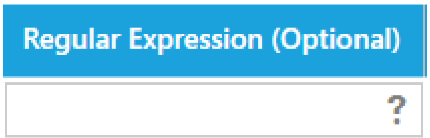
See section 5.5 for more information on tag limits.



Tag Limit

0 + −

d. Optionally, you can specify a regular expression so that only text that match the specified regular expression is used for comparison against the Terms in the Term Store. To find out more about regular expressions click on the ? icon.



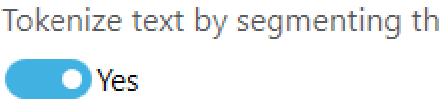
Regular Expression (Optional)

?

e. Click on the **Delete** button to delete any unwanted rows or to start over.

🗑

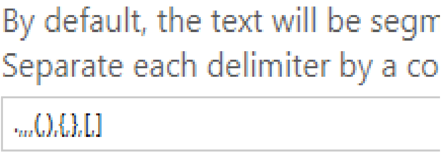
6 Set this to 'Yes' to tokenize the documents' text. This can improve the comparison accuracy. See section 5.2 for more information about Tokenization.



Tokenize text by segmenting th

Yes

7 Set any additional delimiters by separating each one with a comma. The default values should be adequate for most situation.

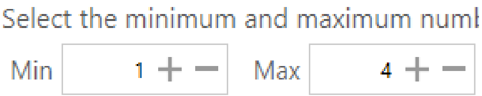


By default, the text will be segm

Separate each delimiter by a co

',,,,(),{},[]

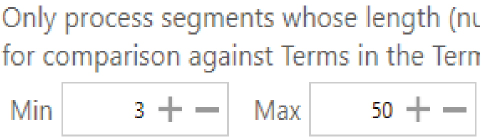
8 Set the number of words that can be in each token.



Select the minimum and maximum numl

Min 1 + − Max 4 + −

9 Set the number of characters that each token must have in order to be used for comparison against Terms. This is useful to avoid comparing words like 'a', 'is', 'to' etc.

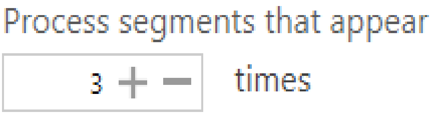


Only process segments whose length (nu

for comparison against Terms in the Tern

Min 3 + − Max 50 + −

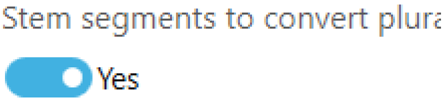
10 Set the minimum frequency a token must appear in a document for it to be used for comparison.



Process segments that appear

3 + − times

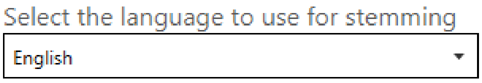
11 Enable stemming to convert plural words to singular to improve comparison accuracy further.



Stem segments to convert plura

Yes

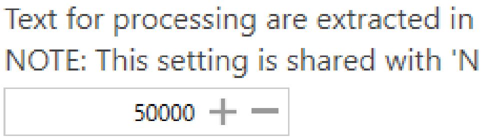
12 Set the language to use for stemming. Different languages have different rules for converting plural to singular.



Select the language to use for stemming

English

13 Each document's text is processed in chunks of 50,000 characters by default. Since this setting is shared with the equivalent NLP setting, make sure it does not interfere with it if you change the default value.



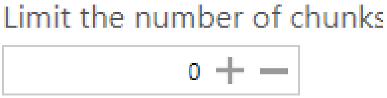
Text for processing are extracted in

NOTE: This setting is shared with 'N

50000 + −

14 Set this to only process the first 'x' number of chunks. This can be useful if you are processing very large documents and the text you want to use for comparison are on the first few pages.

Set '0' to process the whole document.

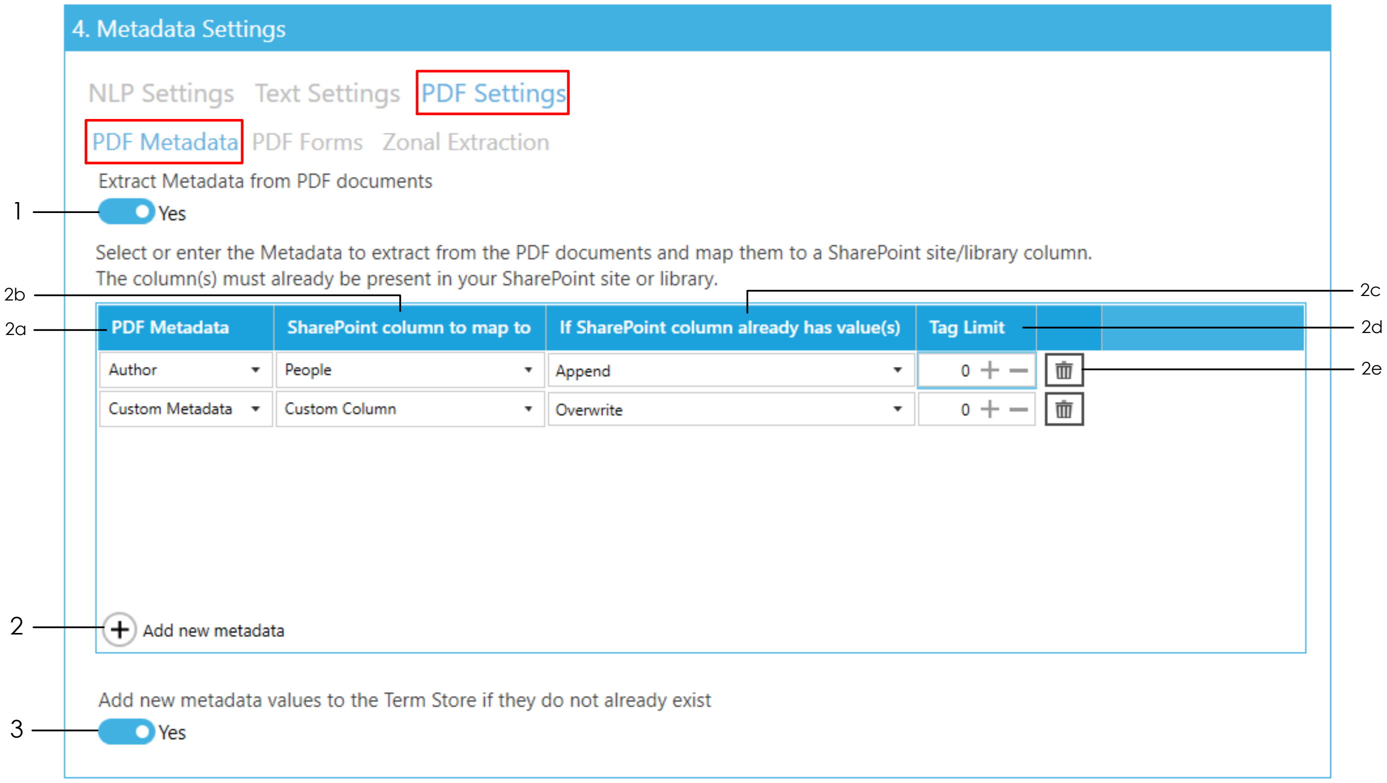


Limit the number of chunks

0 + −

4.5.3 PDF Metadata

This section is used to extract metadata from PDF documents. To access it go to **Job** > **Metadata** > **PDF Settings** > **PDF Metadata** tab.



4. Metadata Settings

NLP Settings Text Settings PDF Settings

PDF Metadata PDF Forms Zonal Extraction

Extract Metadata from PDF documents

Yes

Select or enter the Metadata to extract from the PDF documents and map them to a SharePoint site/library column.
The column(s) must already be present in your SharePoint site or library.

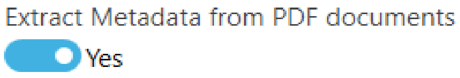
PDF Metadata	SharePoint column to map to	If SharePoint column already has value(s)	Tag Limit
Author	People	Append	0
Custom Metadata	Custom Column	Overwrite	0

Add new metadata

Add new metadata values to the Term Store if they do not already exist

Yes

1 Set **Extract Metadata from PDF documents** to 'Yes'

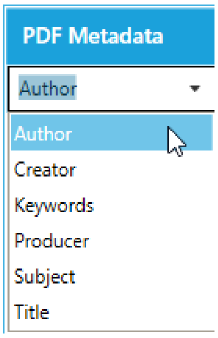


Extract Metadata from PDF documents

Yes

2 To add a PDF Metadata to extract, click on **Add new metadata** button:

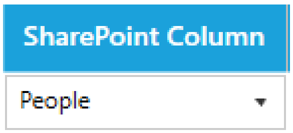
a. Select or type in the name of the PDF metadata to extract. The drop down menu contains a list of common PDF metadata such as 'Title', 'Author', etc. However, you can also add custom metadata by typing them in.



PDF Metadata

Author

b. Select or type in the name of the SharePoint column to add the PDF metadata to. This can be either a Managed Metadata column or a non-Managed Metadata column.



SharePoint Column

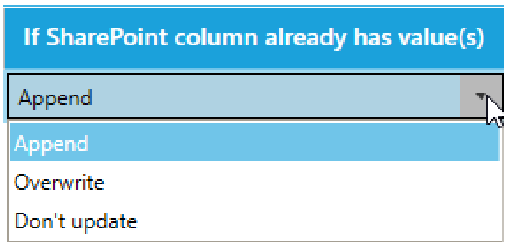
People

c. Select what to do if the SharePoint column you want to add the metadata to already has values:

i. **Append** – the metadata will be appended to the existing values of the SharePoint column

ii. **Overwrite** – the existing values of the of the SharePoint column will be deleted and replaced with the new metadata

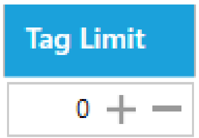
iii. **Don't update** – if the SharePoint column already has values, PDF metadata extraction for this column will be skipped



d. Enter a **Tag Limit**. This restricts the number of extracted metadata that will be added (tagged) to this SharePoint column.

Set '0' for no limits.

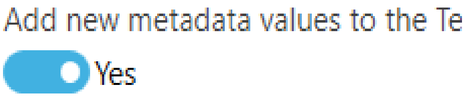
See section 5.5 for more information on tag limits.



e. Click on the **Delete** button to delete any unwanted rows or to start over.

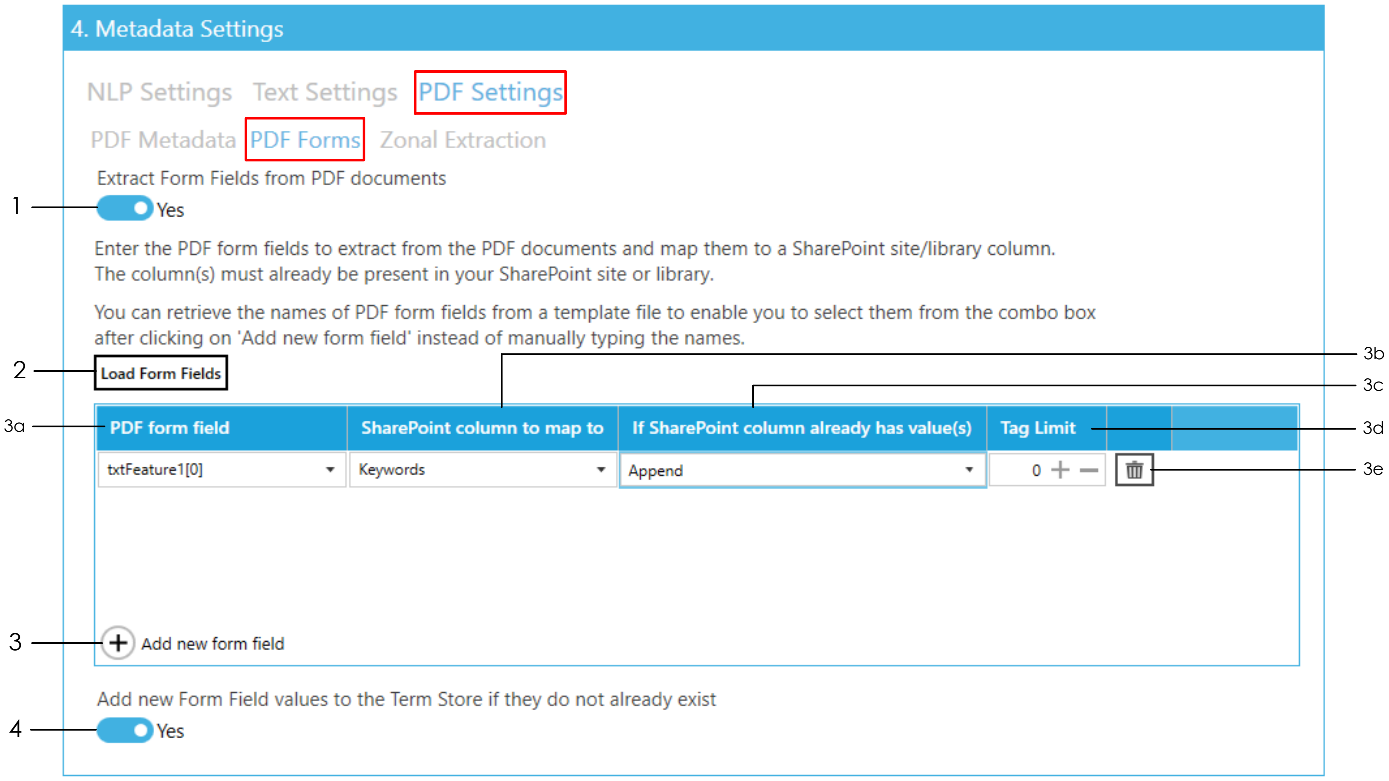
3 If any of the SharePoint columns added in step 2b above is a Managed Metadata column, it will be associated with an existing Term Set in the Term Store.

Set this option to 'Yes' if you want to add metadata extracted from the PDF that are not currently present in the Term Set to the Term Store. Otherwise, the metadata cannot be tagged and will be skipped.



4.5.4 PDF Forms

The following settings enables the extraction of PDF Form Field data from PDF documents so that they can be added to specific SharePoint columns. To access these settings go to **Job** > **Metadata** > **PDF Settings** > **PDF Forms** tab.



4. Metadata Settings

NLP Settings | Text Settings | PDF Settings

PDF Metadata | PDF Forms | Zonal Extraction

Extract Form Fields from PDF documents

Yes

Enter the PDF form fields to extract from the PDF documents and map them to a SharePoint site/library column.
The column(s) must already be present in your SharePoint site or library.

You can retrieve the names of PDF form fields from a template file to enable you to select them from the combo box after clicking on 'Add new form field' instead of manually typing the names.

Load Form Fields

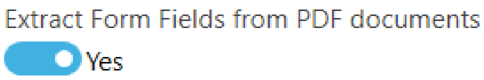
PDF form field	SharePoint column to map to	If SharePoint column already has value(s)	Tag Limit
txtFeature1[0]	Keywords	Append	0

Add new form field

Add new Form Field values to the Term Store if they do not already exist

Yes

1 Set **Extract Form Fields from PDF documents** to 'Yes'



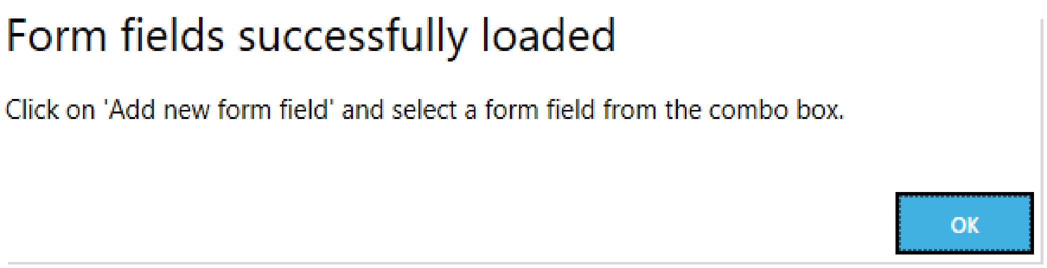
Extract Form Fields from PDF documents

Yes

2 Click on the **Load Form Fields** button and select a template PDF file (sample file) containing the Form Field(s) you want to extract. This will load all the Form Fields from the selected PDF file which will be used in the next step.

Load Form Fields

If the selected file contains Form Fields, you should see the following message:

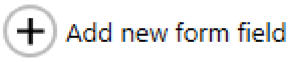


Form fields successfully loaded

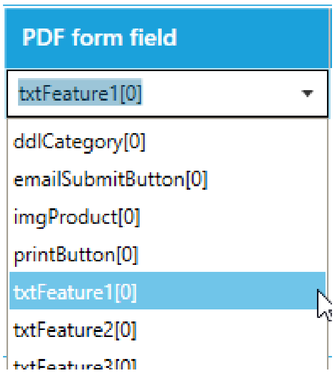
Click on 'Add new form field' and select a form field from the combo box.

OK

3 To add a PDF Form Field data to extract, click on **Add new form field** button:

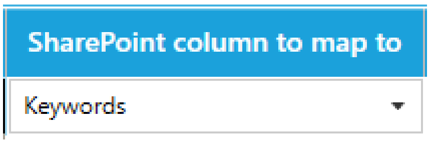


a. Select or type in the name of the PDF Form Field to extract data from. The drop down menu should contain all the Form Fields loaded in step 2. However, you can skip step 2 and type in the name of the Form Field if you already know it.


PDF form field

txtFeature1[0]

b. Select or type in the name of the SharePoint column to add the PDF Form Field data to. This can be either a Managed Metadata column or a non-Managed Metadata column.


SharePoint column to map to

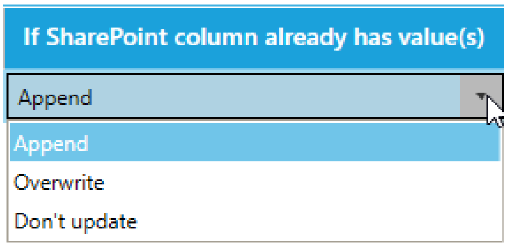
Keywords

c. Select what to do if the SharePoint column you want to add the Form Field data to already has values:

i. **Append** – the Form Field data will be appended to the existing values of the SharePoint column

ii. **Overwrite** – the existing values of the of the SharePoint column will be deleted and replaced with the new Form Field data

iii. **Don't update** – if the SharePoint column already has values, Form Field extraction for this column will be skipped

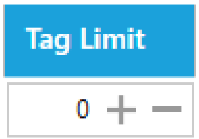

If SharePoint column already has value(s)

Append

d. Enter a **Tag Limit**. This restricts the number of extracted Form Field data that will be added (tagged) to this SharePoint column.

Set '0' for no limits.

See section 5.5 for more information on tag limits.

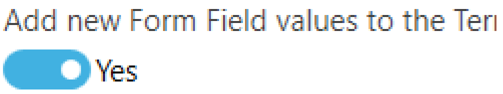

Tag Limit

0

e. Click on the **Delete** button to delete any unwanted rows or to start over.

4 If any of the SharePoint columns added in step 3b above is a Managed Metadata column, it will be associated with an existing Term Set in the Term Store.

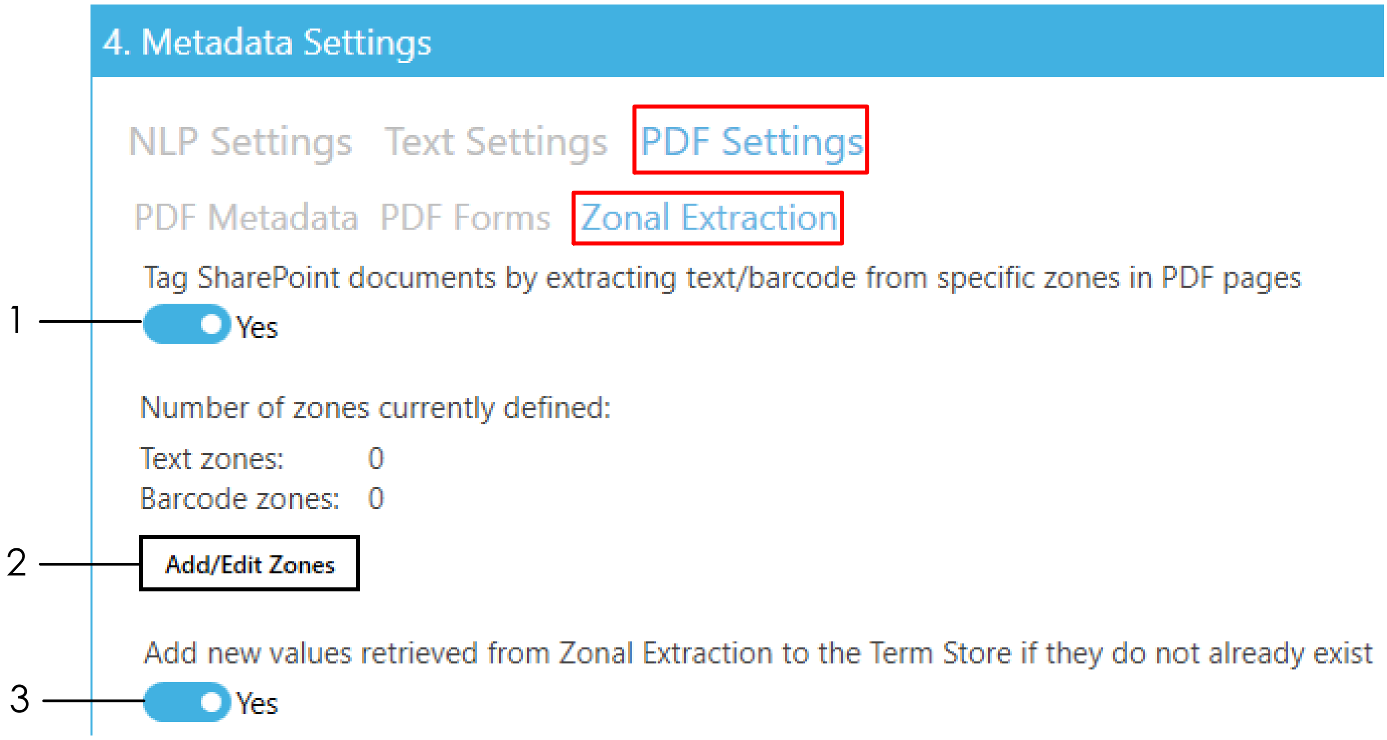
Set this option to 'Yes' if you want to add metadata extracted from the PDF that are not currently present in the Term Set to the Term Store. Otherwise, the metadata cannot be tagged and will be skipped.


Add new Form Field values to the Teri

Yes

4.5.5 Zonal Extraction

To extract text or barcode from specific zones from PDF files and add them to specific SharePoint columns, click on the **Metadata** > **PDF Settings** > **Zonal Extraction** tab.



4. Metadata Settings

Tag SharePoint documents by extracting text/barcode from specific zones in PDF pages

1 Yes

Number of zones currently defined:

Text zones: 0

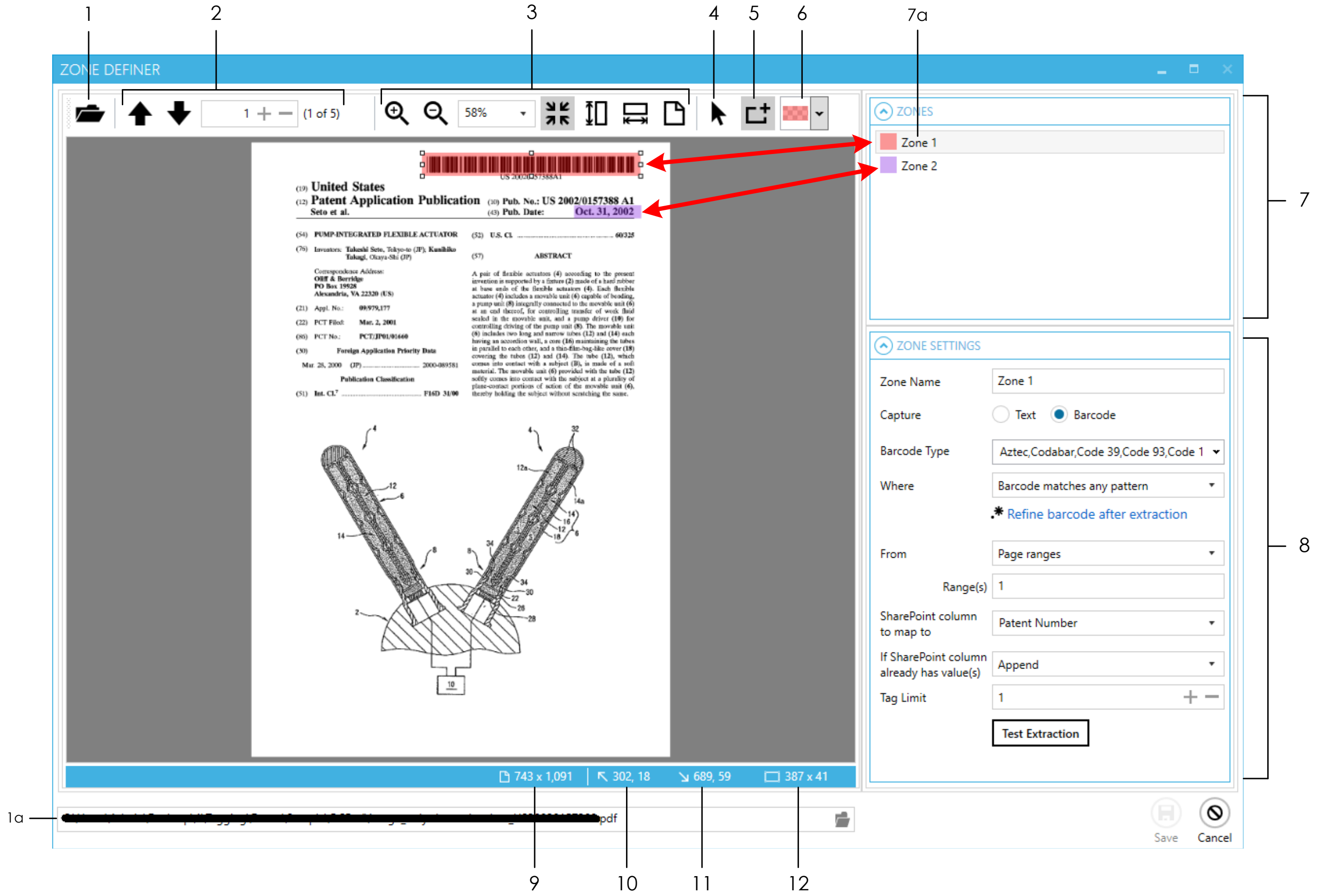
Barcode zones: 0

2 Add/Edit Zones

Add new values retrieved from Zonal Extraction to the Term Store if they do not already exist

3 Yes

4.5.5.1 Zone Definer



ZONE DEFINER

1 + − (1 of 5)   58%

ZONES
- Zone 1
- Zone 2

ZONE SETTINGS

Zone Name: Zone 1
Capture: ○ Text ● Barcode
Barcode Type: Aztec,Codabar,Code 39,Code 93,Code 1
Where: Barcode matches any pattern
Refine barcode after extraction
From: Page ranges
Range(s): 1
SharePoint column to map to: Patent Number
If SharePoint column already has value(s): Append
Tag Limit: 1 + −
Test Extraction

743 x 1,091 | 302, 18 | 689, 59 | 387 x 41

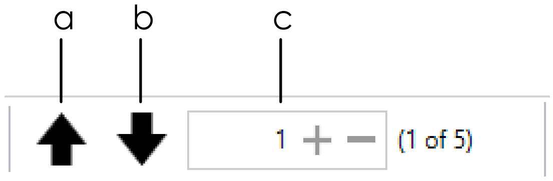
[redacted]pdf

Save Cancel

1 Use this to open the file to use as a template for creating zones

a. Alternatively, use the control at the bottom to perform the same task

2 Navigate to pages in a multipage document



1 + − (1 of 5)

a. Navigates to the previous page

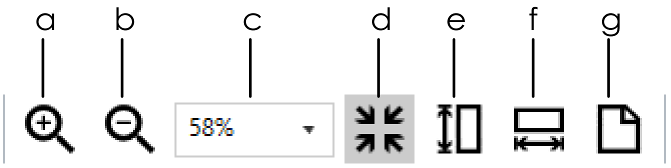
b. Navigates to the next page

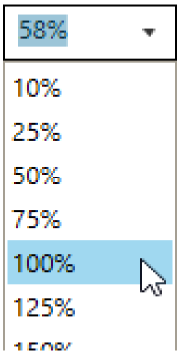
c. Navigates to a specific page by specifying the page number in the text box

1 + − (1 of 5)

Clicking + or − will have the same effect as clicking ⬆ or ⬇ respectively.

3 Use these for zooming pages to the desired size

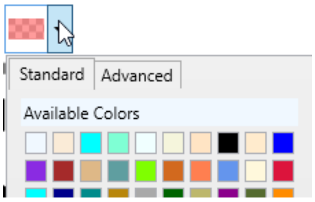


a.	Zoom in to increase magnification of the current page	⊕
b.	Zoom out to decrease magnification of the current page	⊖
c.	Set custom zoom by selecting from the pre-defined zoom levels in the drop-down menu	 58% 10% 25% 50% 75% 100% 125%
d.	Fits the current page to the current size of the Zone Definer window	⤡
e.	Fits the height of the current page to the height of the Zone Definer window	↕
f.	Fits the width of the current page to the width of the Zone Definer window	↔
g.	Zooms to the actual size of the page (100%)	🗋

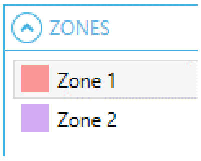
4 Use this to select a zone (that has already been created) from the page ↖

5 Use this to define a new zone

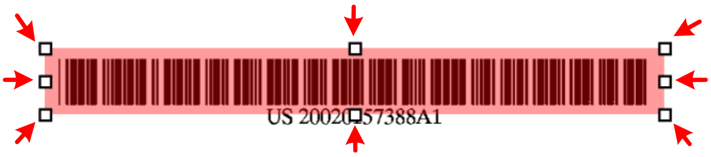
6 Changes the colour of the selected zone. If no zone is currently selected, changing this will affect the colour of the next zone that is created.



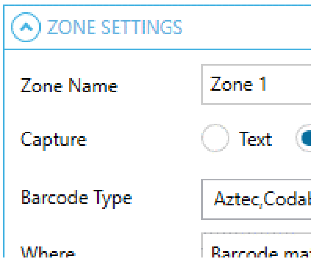
7 The **ZONES** panel shows all the zones currently defined



a. Shows the zone currently selected. Another way to identify which zone is currently selected is to look which zone has the resize controls as indicated by the red arrows below:



8 The **ZONE SETTINGS** panel shows the settings of the zone currently selected



ZONE SETTINGS

Zone Name: Zone 1

Capture: ○ Text ●

Barcode Type: Aztec,Codal

Where: Barcode ma

9 Shows the actual size of the current page: 743 x 1,091

10 The top-left co-ordinates of the selected zone: 302, 18

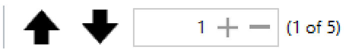
11 The bottom-right co-ordinates of the selected zone: 689, 59

12 The dimension of the selected zone: 387 x 41

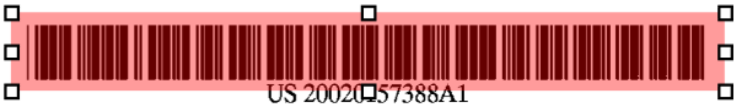
4.5.5.2 Zonal Barcode Extraction

1 On the Zone Definer window, open a PDF file to use as a template for defining the zone to extract the barcode from.

2 Navigate to the page where you want to extract the barcode from.

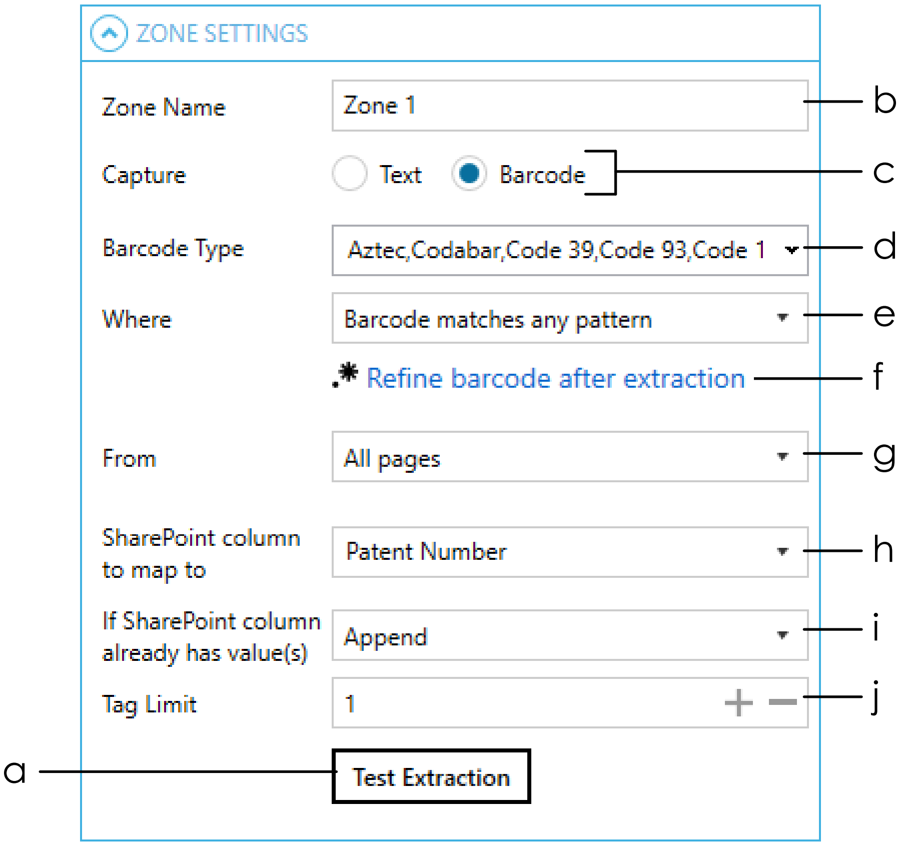


3 From the toolbar at the top, select the Define a zone tool. The cursor should change to a crosshair +. Click and drag on the page to define a zone where the barcode is. You can resize or move the zone after defining it to adjust the size and location as needed.



You can also change the colour of the zone if you want to. This can helpful to differentiate between zones when you have multiple zones defined.

4 With the zone selected, go to the **ZONE SETTINGS** panel to configure its settings

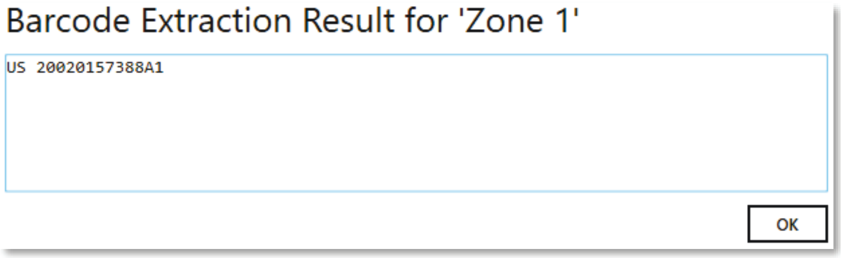


ZONE SETTINGS

Zone Name	Zone 1
Capture	○ Text ● Barcode
Barcode Type	Aztec,Codabar,Code 39,Code 93,Code 1
Where	Barcode matches any pattern
	Refine barcode after extraction
From	All pages
SharePoint column to map to	Patent Number
If SharePoint column already has value(s)	Append
Tag Limit	1

a. Click on the **Test Extraction** button to make sure that the zone currently defined is the correct size and in the correct location.

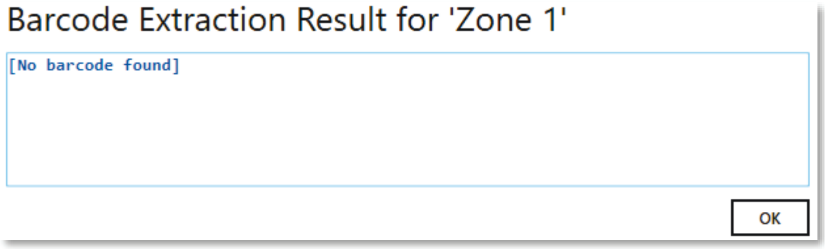
If it is the correct size and in the correct location, it will display the value of the barcode as shown below:



Barcode Extraction Result for 'Zone 1'

```
US 20020157388A1
```

If it is not the correct size or in the wrong location, it will display '[No barcode found]':

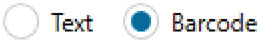


Barcode Extraction Result for 'Zone 1'

```
[No barcode found]
```

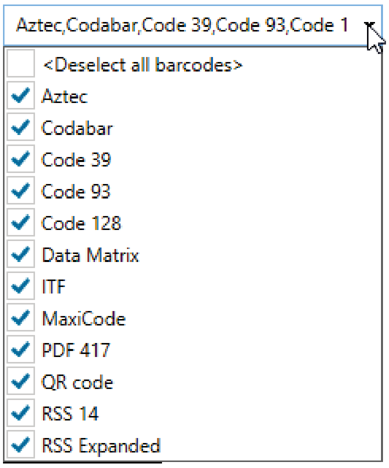
b. Change the default name of the zone if needed

c. Choose **Barcode** as the 'Capture' type



○ Text ● Barcode

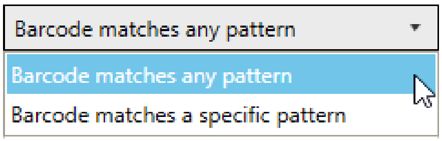
d. Select the **Barcode Type** that the SharePoint documents you want to extract barcodes from have. Select all if you do not know the barcode type(s).



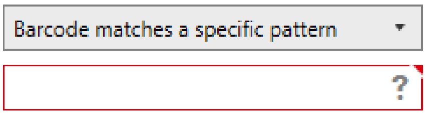
Aztec,Codabar,Code 39,Code 93,Code 1

- □ <Deselect all barcodes>
- ☑ Aztec
- ☑ Codabar
- ☑ Code 39
- ☑ Code 93
- ☑ Code 128
- ☑ Data Matrix
- ☑ ITF
- ☑ MaxiCode
- ☑ PDF 417
- ☑ QR code
- ☑ RSS 14
- ☑ RSS Expanded

e. Select whether you want the extracted barcode to match a specific pattern or not


Barcode matches any pattern
Barcode matches any pattern
Barcode matches a specific pattern

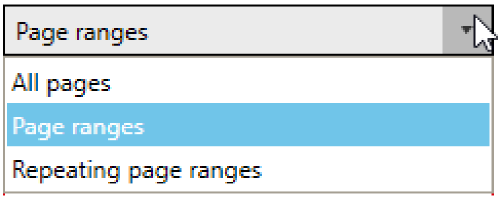
If you select "**Barcode matches a specific pattern**", a textbox will appear at the bottom for you to specify the pattern (regular expression).



f. See Post Extraction Settings for explanation of this feature.

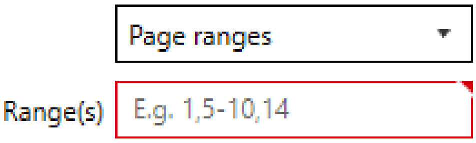
g. Choose which page(s) to extract the barcode from.

i. **All Pages –** Tagger will attempt to extract barcode from all pages at the specified zone



ii. **Page Ranges –** Tagger will attempt to extract barcode only from the specified page ranges.

If you select this option, a textbox will appear at the bottom for you to specify the page(s). Pages and page ranges must be separated by a comma.



Example:

1	Page 1 only
3-6	Pages 3,4,5,6
1,3-6,14	Pages 1,3,4,5,6,14

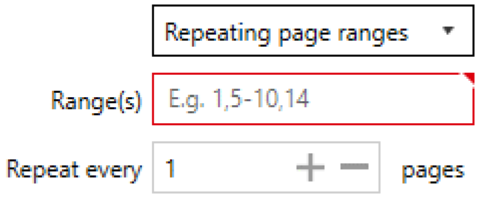
iii. **Repeating page ranges –** Similar to **Page Ranges** but with a different method of specifying the pages to extract the barcode from.

With this option selected, you will be provided with 2 additional textboxes to fill.

- The **Range(s)** textbox is same as described above in ii.

- The **Repeat Every** value specifies the interval to re-apply the page range specified.

E.g. if **Range(s)** is set to 3-6 and **Repeat Every** is set to 5, the range is re-applied every 5 pages from the starting page “3”, hence resulting in the following pages: 3,4,5,6 then 8,9,10,11 then 13,14,15,16 and so on.

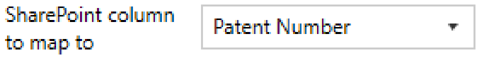


Repeating page ranges

Range(s) E.g. 1,5-10,14

Repeat every 1 pages

h. Select or type in the name of the SharePoint column to add the extracted barcode to. This can be either a Managed Metadata column or a non-Managed Metadata column.



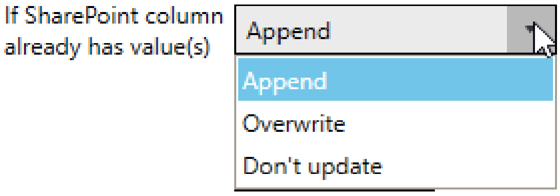
SharePoint column to map to: Patent Number

i. Select what to do if the SharePoint column you want to add the barcode to already has values:

i. **Append** – the barcode will be appended to the existing values of the SharePoint column

ii. **Overwrite** – the existing values of the of the SharePoint column will be deleted and replaced with the new barcode

iii. **Don’t update** – if the SharePoint column already has values, the barcode extraction for this column will be skipped

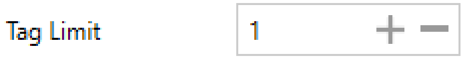


If SharePoint column already has value(s): Append

j. Enter a **Tag Limit**. This restricts the number of extracted barcodes that will be added (tagged) to this SharePoint column.

Set ‘0’ for no limits.

See section 5.5 for more information on tag limits.



Tag Limit 1

5. Click **Save** at the bottom of the Zone Definer if you don’t have any more zones to define. Otherwise repeat from step 2 above to create another zone to extract barcode or go to the next section for steps in defining zones to extract text.

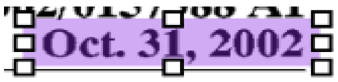
Save

4.5.5.3 Zonal Text Extraction

1 On the Zone Definer window, open a PDF file to use as a template for defining the zone to extract the text from.

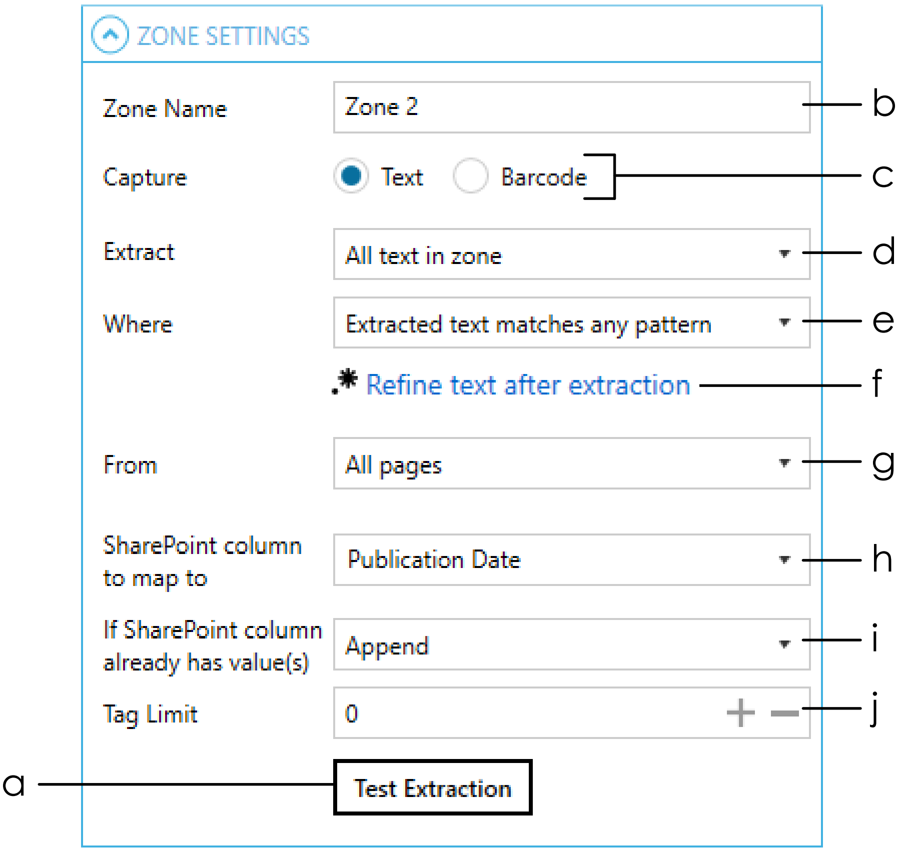
2 Navigate to the page where you want to extract the text from.

3 From the toolbar at the top, select the Define a zone tool. The cursor should change to a crosshair +. Click and drag on the page to define a zone where the text you want to extract is. You can resize or move the zone after defining it to adjust the size and location as needed.



You can also change the colour of the zone if you want to. This can helpful to differentiate between zones when you have multiple zones defined.

4 With the zone selected, go to the **ZONE SETTINGS** panel to configure its settings

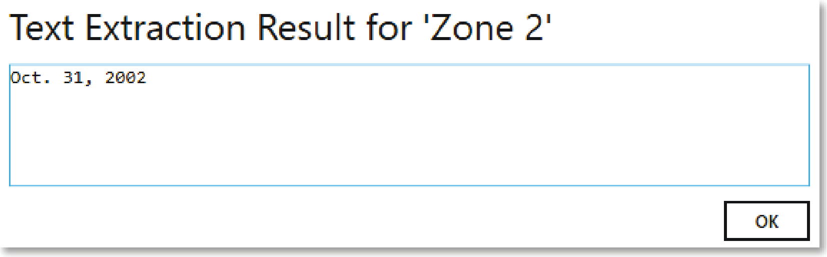


ZONE SETTINGS

Zone Name	Zone 2
Capture	● Text ○ Barcode
Extract	All text in zone
Where	Extracted text matches any pattern
	Refine text after extraction
From	All pages
SharePoint column to map to	Publication Date
If SharePoint column already has value(s)	Append
Tag Limit	0

a. Click on the **Test Extraction** button to make sure that the zone currently defined is the correct size and in the correct location.

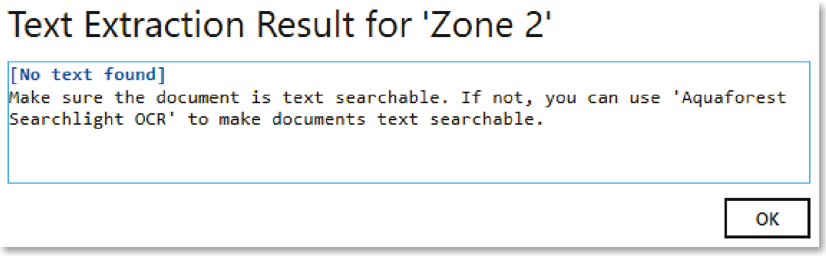
If it is the correct size and in the correct location, it will display the value of the text as shown below:



Text Extraction Result for 'Zone 2'

```
Oct. 31, 2002
```

If it is not the correct size or in the wrong location, it will display '[No text found]':



Text Extraction Result for 'Zone 2'

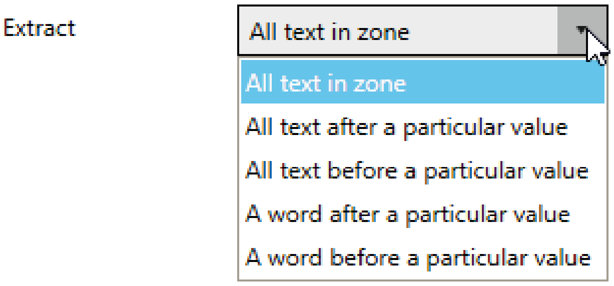
```
[No text found]
Make sure the document is text searchable. If not, you can use 'Aquaforest
Searchlight OCR' to make documents text searchable.
```

b. Change the default name of the zone if needed

c. Choose **Text** as the 'Capture' type

● Text ○ Barcode

d. Select how you want to extract the text from the specified zone. The following options are available:



Extract: All text in zone

All text in zone
All text after a particular value
All text before a particular value
A word after a particular value
A word before a particular value

i. **All text in zone –** this is useful if all documents are exactly the same in terms of page size and content structure such that the zone will always point to the same location in all documents.

ii. **All text after a particular value**

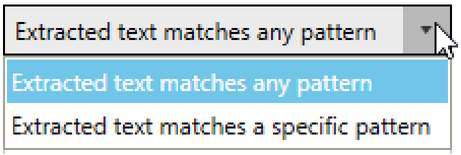
iii. **All text before a particular value**

iv. **A word after a particular value**

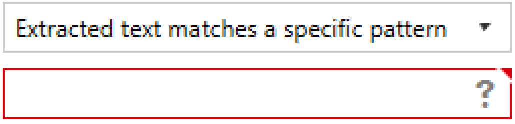
v. **A word before a particular value**

Options ii, iii, iv and v are useful if the documents may not have the same size pages or its content structure may be different but the text you want to extract always comes before or after a particular value such as a field name like “Invoice no.”.

e. Select whether you want the extracted text to match a specific pattern or not


Extracted text matches any pattern
Extracted text matches any pattern
Extracted text matches a specific pattern

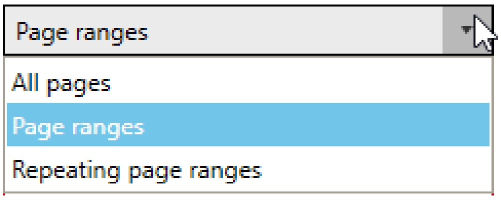
If you select “**Extracted text matches a specific pattern**”, a textbox will appear at the bottom for you to specify the pattern (regular expression).



f. See Post Extraction Settings for explanation of this feature.

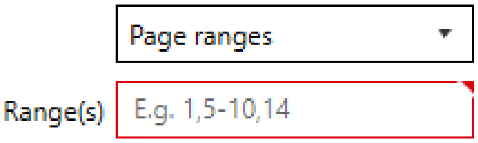
g. Choose which page(s) to extract the barcode from.

vi. **All Pages –** Tagger will attempt to extract text from all pages at the specified zone



vii. **Page Ranges –** Tagger will attempt to extract text only from the specified page ranges.

If you select this option, a textbox will appear at the bottom for you to specify the page(s). Pages and page ranges must be separated by a comma.



Example:

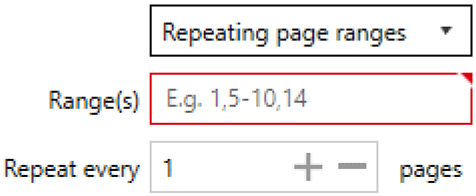
1	Page 1 only
3-6	Pages 3,4,5,6
1,3-6,14	Pages 1,3,4,5,6,14

viii. **Repeating page ranges** – Similar to **Page Ranges** but with a different method of specifying the pages to extract the text from.

With this option selected, you will be provided with 2 additional textboxes to fill.
- The **Range(s)** textbox is same as described above in ii.

- The **Repeat Every** value specifies the interval to re-apply the page range specified.

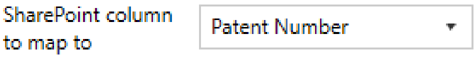
E.g. if **Range(s)** is set to 3-6 and **Repeat Every** is set to 5, the range is re-applied every 5 pages from the starting page "3", hence resulting in the following pages:
3,4,5,6 then 8,9,10,11 then 13,14,15,16 and so on.


Repeating page ranges

Range(s) E.g. 1,5-10,14

Repeat every 1 pages

h. Select or type in the name of the SharePoint column to add the extracted barcode to. This can be either a Managed Metadata column or a non-Managed Metadata column.

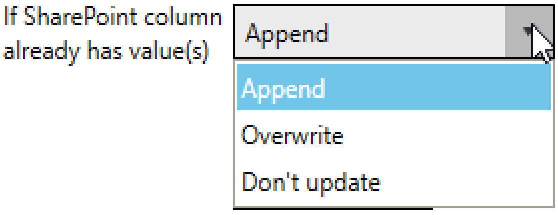

SharePoint column to map to: Patent Number

i. Select what to do if the SharePoint column you want to add the barcode to already has values:

iv. **Append** – the barcode will be appended to the existing values of the SharePoint column

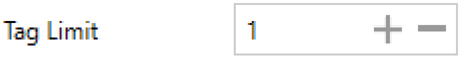
v. **Overwrite** – the existing values of the of the SharePoint column will be deleted and replaced with the new barcode

vi. **Don't update** – if the SharePoint column already has values, the barcode extraction for this column will be skipped


If SharePoint column already has value(s): Append

j. Enter a **Tag Limit**. This restricts the number of extracted barcodes that will be added (tagged) to this SharePoint column.

Set '0' for no limits.


Tag Limit 1

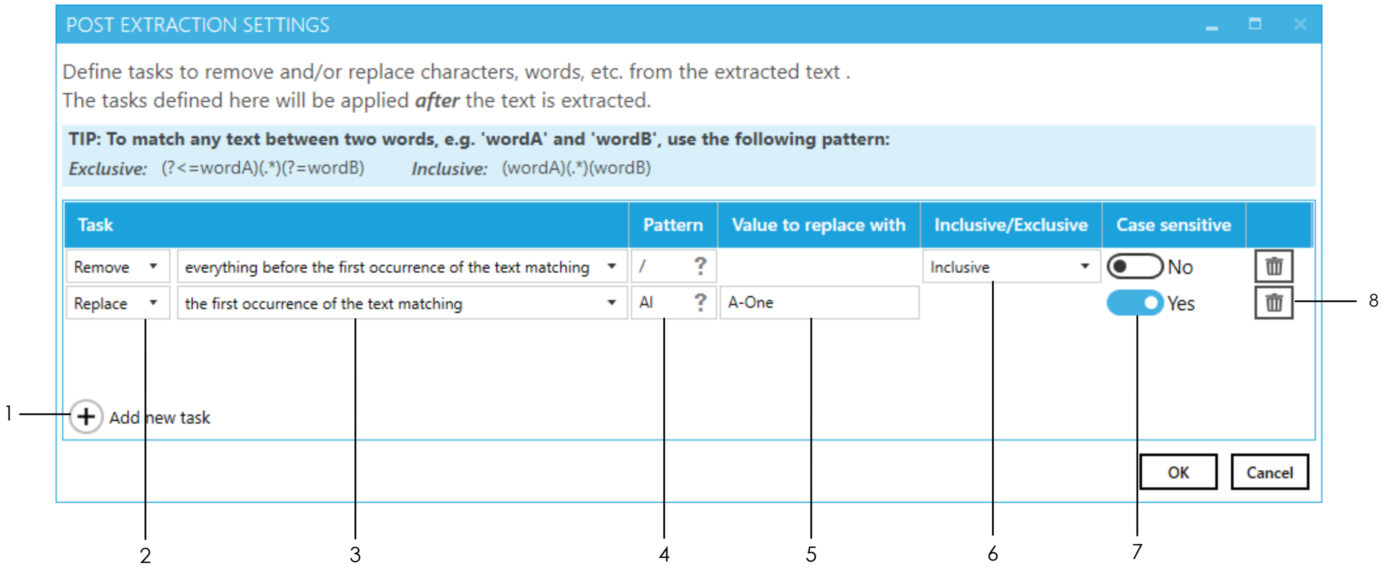
See section 5.5 for more information on tag limits.

5. Click **Save** at the bottom of the Zone Definer if you don't have any more zones to define. Otherwise repeat from step 2 above to create another zone to extract text or go to the previous section for steps in defining zones to extract barcode.

4.5.5.4 Post Extraction Settings

Post Extraction Settings enables you to refine text or barcode further after they have been extracted. It allows removing and/or replacing specific characters, words, numbers, etc. from the extracted text/barcode.

You can access the post extraction settings by clicking the Refine text after extraction or the Refine barcode after extraction link in the Zone Settings panel in the Zone Definer window.



POST EXTRACTION SETTINGS

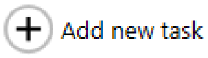
Define tasks to remove and/or replace characters, words, etc. from the extracted text .
The tasks defined here will be applied *after* the text is extracted.

TIP: To match any text between two words, e.g. 'wordA' and 'wordB', use the following pattern:
Exclusive: (?<=wordA)(.*)(?=wordB) *Inclusive:* (wordA)(.*)(wordB)

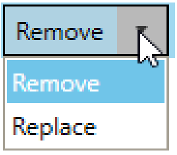
Task		Pattern	Value to replace with	Inclusive/Exclusive	Case sensitive	
Remove	everything before the first occurrence of the text matching	/ ?		Inclusive	No	
Replace	the first occurrence of the text matching	Al ?	A-One		Yes	

To create a Remove or Replace task, follow the following steps:

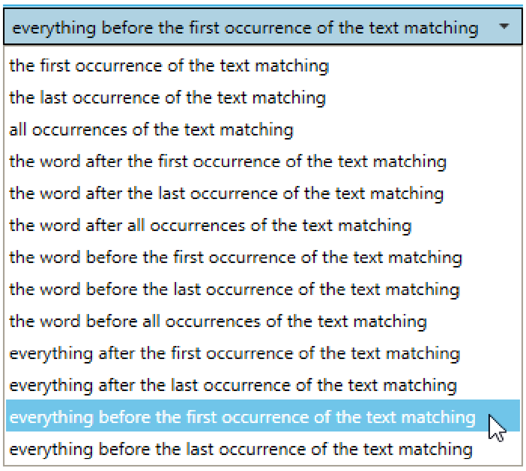
1 Click on the **Add new task** button



2 Select whether you want to **Remove** or **Replace** a particular text from the extracted text or barcode



3 Select the method you want to use to identify the text to be removed or replaced

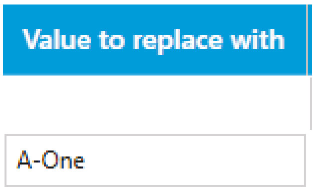


everything before the first occurrence of the text matching
the first occurrence of the text matching
the last occurrence of the text matching
all occurrences of the text matching
the word after the first occurrence of the text matching
the word after the last occurrence of the text matching
the word after all occurrences of the text matching
the word before the first occurrence of the text matching
the word before the last occurrence of the text matching
the word before all occurrences of the text matching
everything after the first occurrence of the text matching
everything after the last occurrence of the text matching
everything before the first occurrence of the text matching
everything before the last occurrence of the text matching

4 Specify a pattern (regular expression) to search the text, character, etc. you want.

5 Specify the value you want to replace the text identified above with.

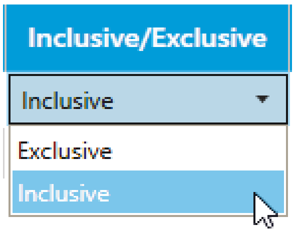
This option is only available if you are create a **Replace** task



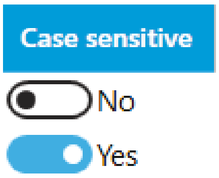
6 Specify whether you want the pattern specified in step 4 to be part of the removal/replacement process or not

i. **Exclusive** – the pattern will *not* be removed or replaced
ii. **Inclusive** – the pattern will be removed or replaced

This option is only available for methods (step 3 above) that contain the word “**before**” or “**after**”.



7 Specify whether the pattern matching should be case-sensitive or not.

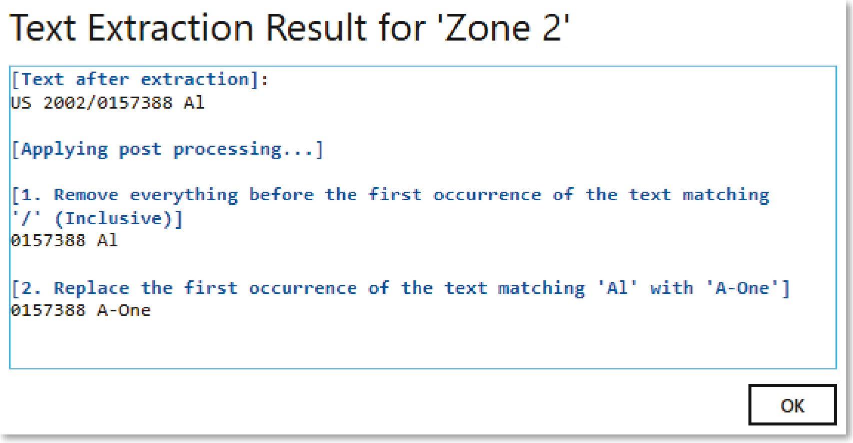


8 Click on the **Delete** button to delete any unwanted rows or to start over. 🗑

Repeat these steps to create more tasks and once finished, click on the **Save** button at the bottom.

To check if the tasks achieve what they are supposed to do, click on the Test Extraction button in the Zone Settings panel in the Zone Definer window.

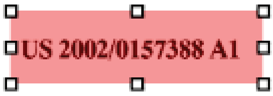
You should see something similar to the following image where it shows step-by-step the text after extraction and the result of applying the post extraction tasks to the extracted text.



Text Extraction Result for 'Zone 2'

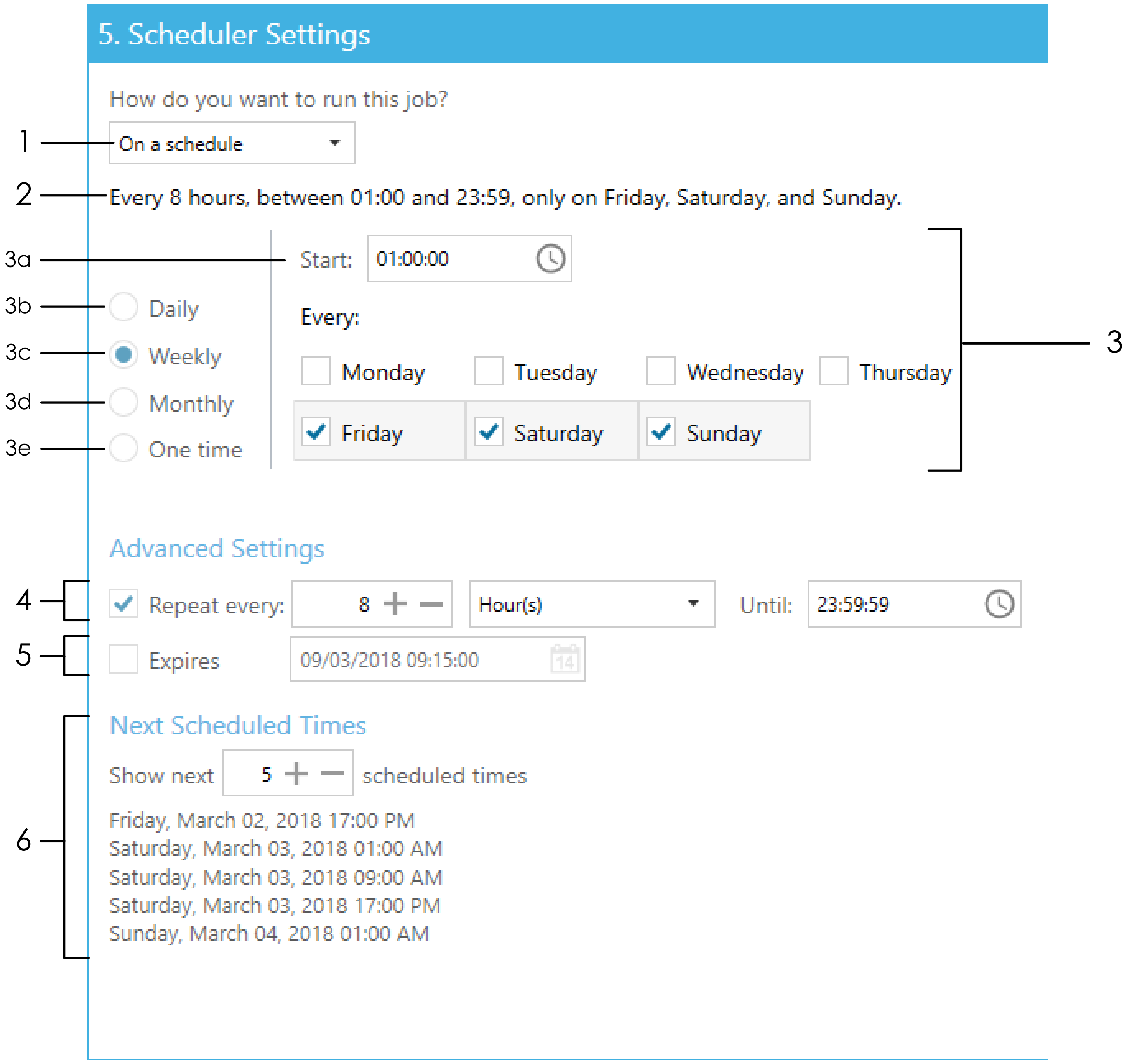
```
[Text after extraction]:
US 2002/0157388 Al

[Applying post processing...]

[1. Remove everything before the first occurrence of the text matching
'/' (Inclusive)]
0157388 Al

[2. Replace the first occurrence of the text matching 'Al' with 'A-One']
0157388 A-One
```

The above Post Extraction tasks were applied to the following zone:



4.6 Scheduler

The scheduler allows you to run jobs automatically at specific times daily, weekly, monthly or on a specific date. Most people choose to run their jobs during off peak hours to avoid users inadvertently locking documents that Tagger may need for processing and to avoid negatively affecting their SharePoint performance.



5. Scheduler Settings

How do you want to run this job?

On a schedule ▾

Every 8 hours, between 01:00 and 23:59, only on Friday, Saturday, and Sunday.

Start: 01:00:00

○ Daily
● Weekly
○ Monthly
○ One time

Every:

□ Monday □ Tuesday □ Wednesday □ Thursday
☑ Friday ☑ Saturday ☑ Sunday

Advanced Settings

☑ Repeat every: 8 + − Hour(s) ▾ Until: 23:59:59

□ Expires 09/03/2018 09:15:00

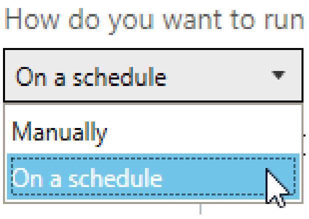
Next Scheduled Times

Show next 5 + − scheduled times

Friday, March 02, 2018 17:00 PM
Saturday, March 03, 2018 01:00 AM
Saturday, March 03, 2018 09:00 AM
Saturday, March 03, 2018 17:00 PM
Sunday, March 04, 2018 01:00 AM

To run the job on a schedule:

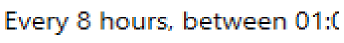
1 Set **How do you want to run this job?** to **On a Schedule**



How do you want to run

On a schedule ▾

2 This is a human readable description of the schedule settings selected. It is automatically updated whenever you make a change to any of the settings described below.



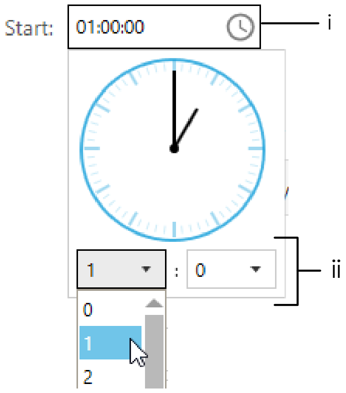
Every 8 hours, between 01:(

3 Choose when you want to run the job.

a. Select a **Start** time. Either

i. type in the time directly, or

ii. select the time using the drop down menus

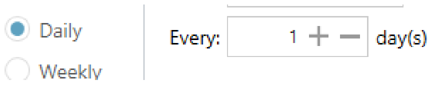


Start: 01:00:00

1 : 0

0
1
2

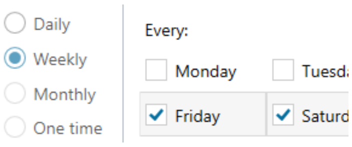
b. **Daily** – Select this if you want to run the job every day or every ‘x’ number of days at the time specified above. Set the value for ‘x’ in the **Every** field.



● Daily
○ Weekly

Every: 1 + — day(s)

c. **Weekly** – Select this if you want to run the job weekly on specific days. Choose which days to run the job by checking the appropriate checkboxes under the **Every** field.

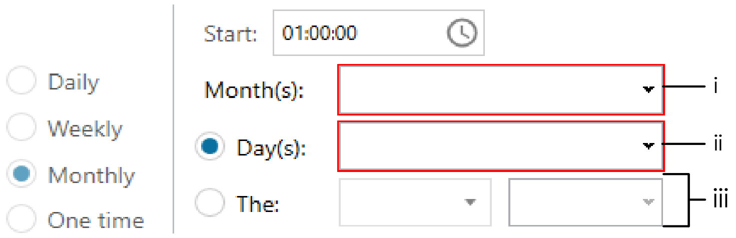


○ Daily
● Weekly
○ Monthly
○ One time

Every:

□ Monday □ Tuesd
☑ Friday ☑ Saturd

d. **Monthly** – Choose this if you want to run the job monthly on specific dates or days of the week



○ Daily
○ Weekly
● Monthly
○ One time

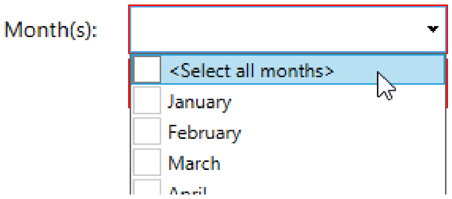
Start: 01:00:00

Month(s):

● Day(s):

○ The:

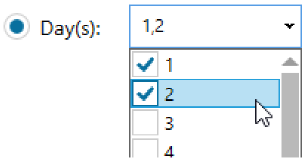
i. Select which month(s) you want to run the job. Choose **<Select all months>** to run the job each month



Month(s):

□ <Select all months>
□ January
□ February
□ March
□ April

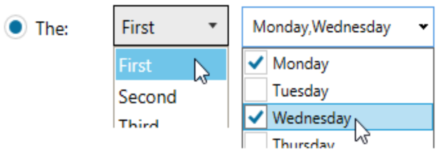
ii. *Either*, select which date(s) of the selected month(s) you want to run the job



● Day(s): 1,2

☑ 1
☑ 2
□ 3
□ 4

iii. *Or*, select whether you want to run the job on the first, second, third, fourth of a particular week day of the selected month(s).

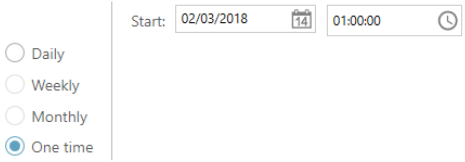


● The: First ▾ Monday,Wednesday ▾

First
Second
Third

☑ Monday
☐ Tuesday
☑ Wednesday
☐ Thursday

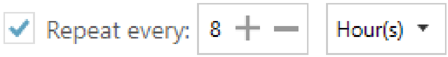
e. One time – select this if you want to run the job only once on a specific date and time



Start: 02/03/2018 01:00:00

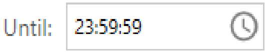
○ Daily
○ Weekly
○ Monthly
● One time

4 Choose whether you want to repeat running the job every ‘x’ minutes or hours after the specified date(s)/time(s) above.



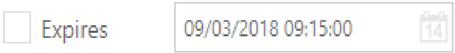
☑ Repeat every: 8 + − Hour(s) ▾

You can also specify until what time to repeat.



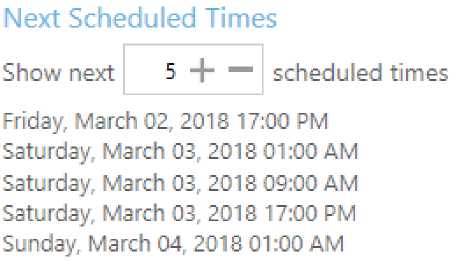
Until: 23:59:59

5 Specify whether you want the schedule to expire or not.



☐ Expires 09/03/2018 09:15:00

6 Update this to view the next ‘x’ scheduled times (from the current time) that the job will run based on the scheduler settings specified.



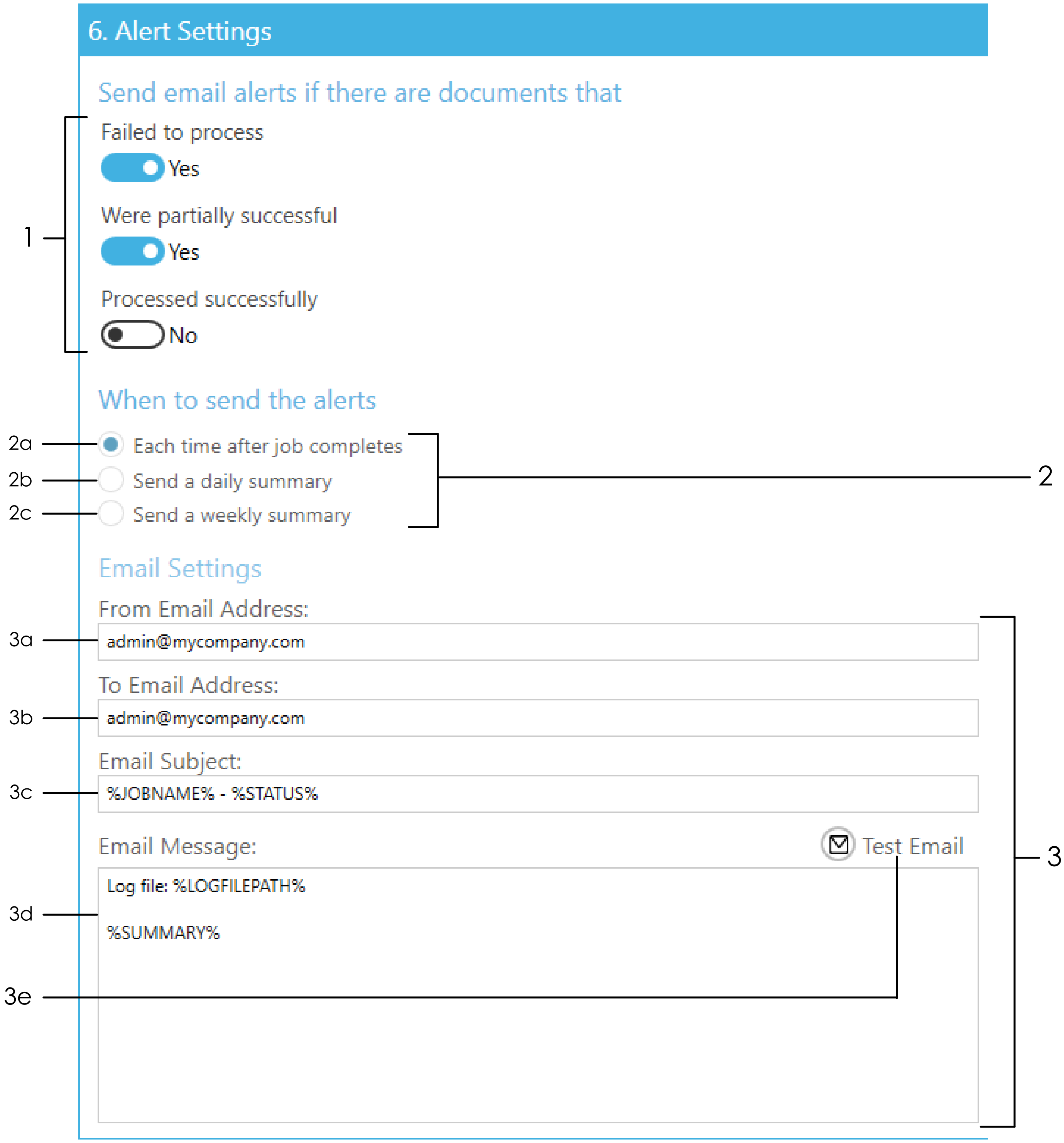
Next Scheduled Times

Show next 5 + − scheduled times

Friday, March 02, 2018 17:00 PM
Saturday, March 03, 2018 01:00 AM
Saturday, March 03, 2018 09:00 AM
Saturday, March 03, 2018 17:00 PM
Sunday, March 04, 2018 01:00 AM

4.7 Alerts

Alerts notifies you by email when a job was successful, partially successful and/or failed. To configure alerts, you must first configure your Email SMPTP settings, or else the email cannot be sent. See section 4.11 for configuring SMTP settings.



6. Alert Settings

Send email alerts if there are documents that

Failed to process

Yes

Were partially successful

Yes

Processed successfully

No

When to send the alerts

● Each time after job completes

○ Send a daily summary

○ Send a weekly summary

Email Settings

From Email Address: admin@mycompany.com

To Email Address: admin@mycompany.com

Email Subject: %JOBNAME% - %STATUS%

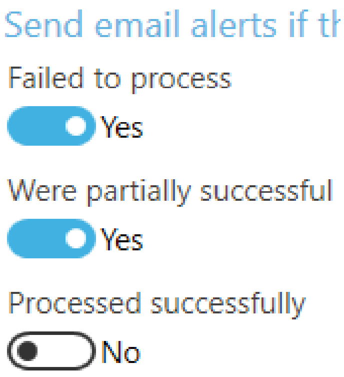
Email Message:

Log file: %LOGFILEPATH%

%SUMMARY%

1 Select whether you want alerts to be sent if the job has documents that:

i. **Failed to process**
ii. **Were partially successful**
iii. **Processed successfully**


Send email alerts if tl

Failed to process
Yes

Were partially successful
Yes

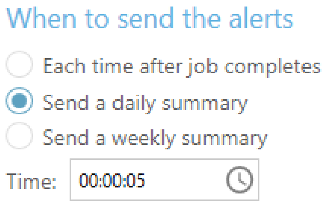
Processed successfully
No

2 Choose when to send the alerts.

a. Each time after job completes

b. Send a daily summary

Specify the time to send the alerts

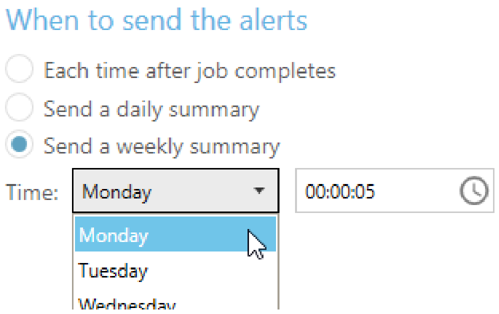

When to send the alerts

○ Each time after job completes
● Send a daily summary
○ Send a weekly summary

Time: 00:00:05

c. Send a weekly summary

Specify the day and time to send the alerts

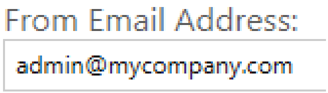

When to send the alerts

○ Each time after job completes
○ Send a daily summary
● Send a weekly summary

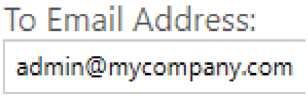
Time: Monday 00:00:05

3 Specify the email settings

a. The email address to send the email from

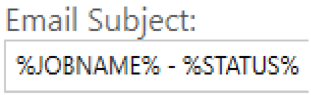

From Email Address: admin@mycompany.com

b. The email address where the alerts will be sent to


To Email Address: admin@mycompany.com

c. The email subject. You can use the following templates:

i. ***%JOBNAME%*** - will be replaced by the name of the library
ii. ***%STATUS%*** - if **'Each time after job completes**' is selected, this will be replaced by status of the job

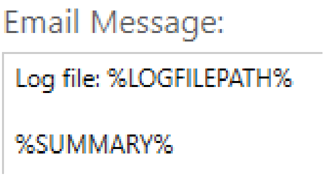

Email Subject: %JOBNAME% - %STATUS%

d. Type in a template for the **Email Message** to be sent. You can use the following templates:

i. ***%JOBNAME%*** - will be replaced by the name of the library

ii. ***%STATUS%*** - if **'Each time after job completes**' is selected, this will be replaced by status of the job

iii. ***%SUMMARY%*** - will be replaced by a summary of the job(s)



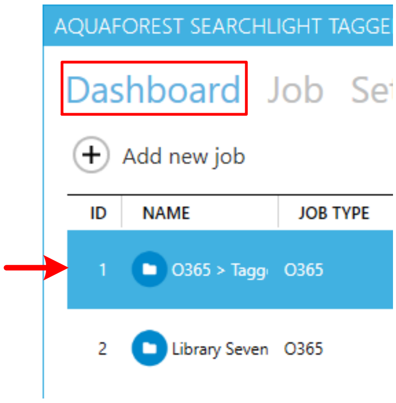
Log file: %LOGFILEPATH%

%SUMMARY%

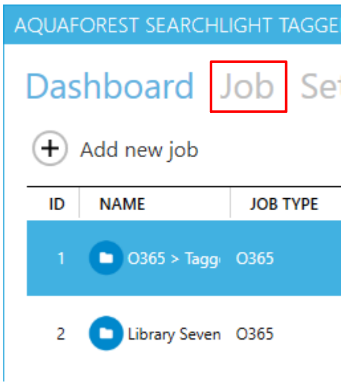
e. After specifying the settings above, you can test the alert by clicking the **Test Email** button.

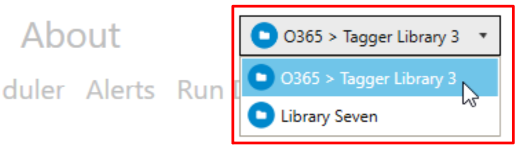
4.8 Editing a Job

1 Select the job from the **Dashboard**



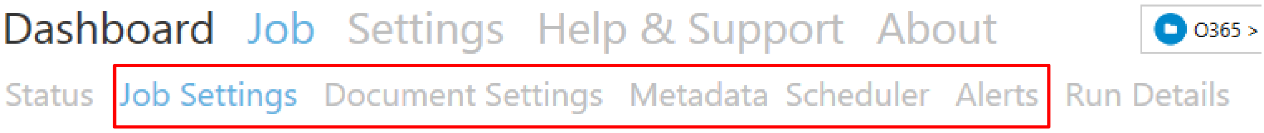
2 Double click on the selected Job or click on the **Job** tab



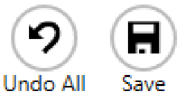


You can also select a job to edit by choosing the library from the combo box at the top of the window.

3 Click on the tab that contain the setting you want to edit.


Dashboard Job Settings Help & Support About
O365 >
Status Job Settings Document Settings Metadata Scheduler Alerts Run Details

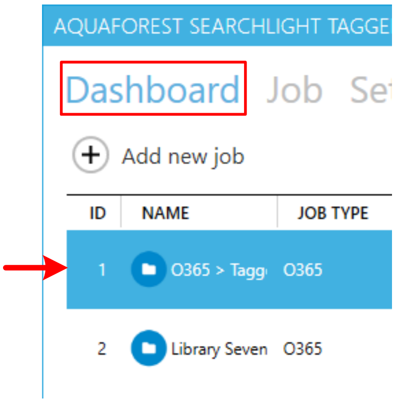
4 Update the setting(s) and click on **Save** at the bottom of the page (you may need to scroll down to see the **Save** button).



Click on the **Undo All** button to undo all changes made.

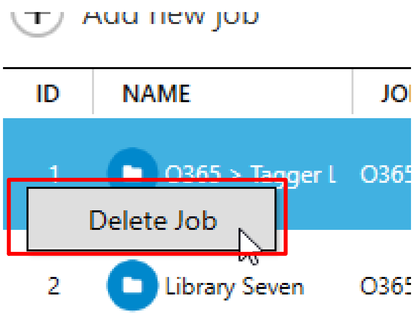
4.9 Deleting a Job

1 Select the job from the **Dashboard**



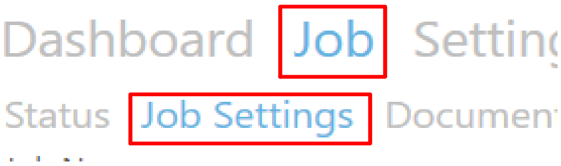
ID	NAME	JOB TYPE
1	O365 > Tagg	O365
2	Library Seven	O365

2 Right click on it and select **Delete Job**

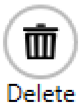


ID	NAME	JO
1	O365 > Tagger L	O365
2	Library Seven	O365

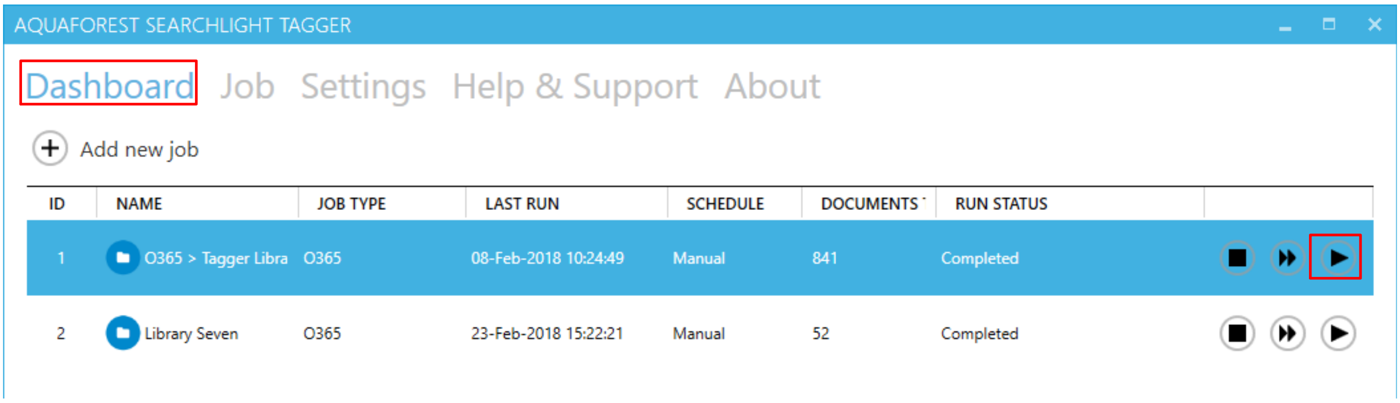
Alternatively, go to **Job** > **Job Settings** tab



Click on the **Delete** button at the bottom of the page



4.10 Running a Job

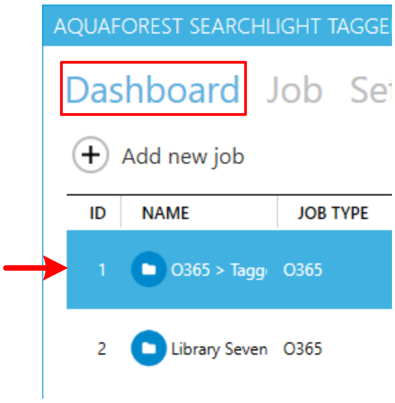

AQUAFOREST SEARCHLIGHT TAGGER

Dashboard Job Settings Help & Support About

Add new job

ID	NAME	JOB TYPE	LAST RUN	SCHEDULE	DOCUMENTS	RUN STATUS
1	O365 > Tagger Libra	O365	08-Feb-2018 10:24:49	Manual	841	Completed
2	Library Seven	O365	23-Feb-2018 15:22:21	Manual	52	Completed

1 Select the job from the **Dashboard**


AQUAFOREST SEARCHLIGHT TAGGE

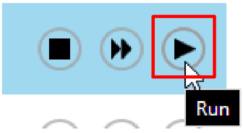
Dashboard Job Se

Add new job

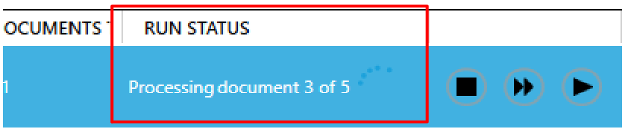
ID	NAME	JOB TYPE
1	O365 > Tagg	O365
2	Library Seven	O365

2 Click on the **Run** button

NOTE: Make sure the service is running. Otherwise, you cannot run a job.

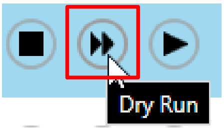


The **Run Status** column should change to "Processing x of y documents"


OCUMENTS RUN STATUS

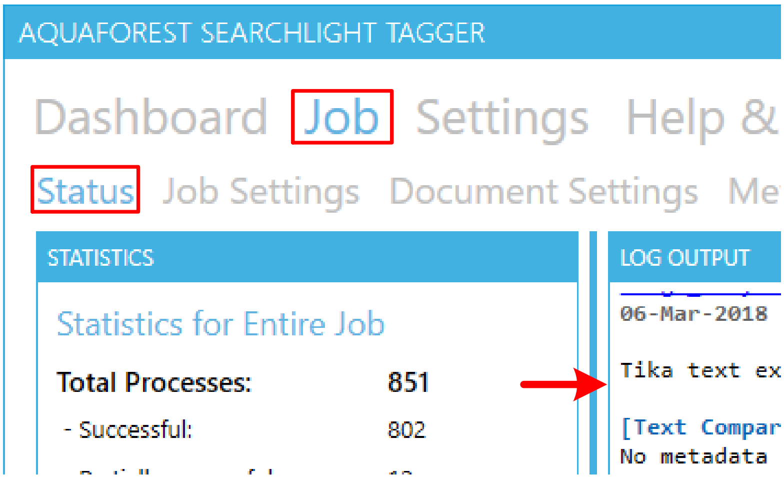
Processing document 3 of 5

If you want to test the settings of the Job without actually updating any documents in SharePoint, you can choose the **Dry Run** button instead.



3 You can view the details of the documents already processed by going to **Job** > **Status** to view the **LOG OUTPUT**.

See section 4.10.1 for more details on the log output.



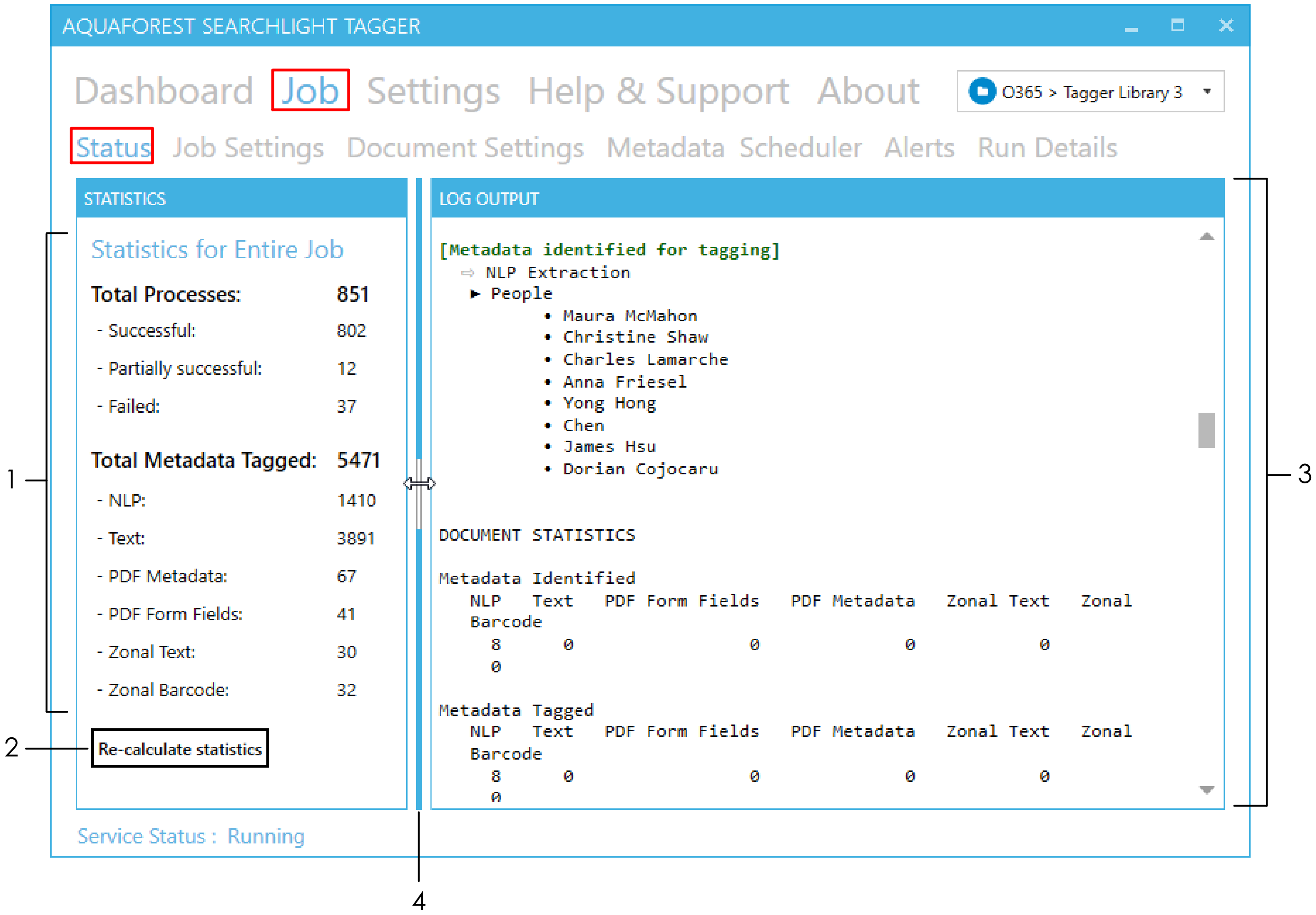
Statistics for Entire Job

Total Processes: 851

- Successful: 802

```
06-Mar-2018
Tika text ex
[Text Compar
No metadata
```

4.10.1 Log Output (Status)



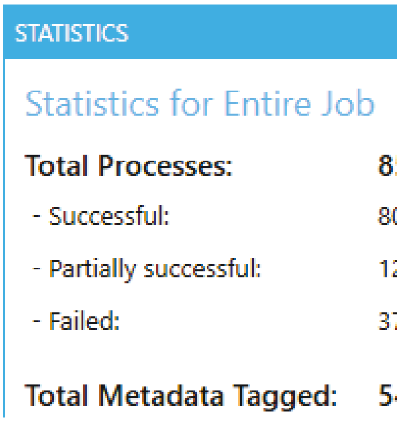
Statistics for Entire Job

Total Processes:	851
- Successful:	802
- Partially successful:	12
- Failed:	37
Total Metadata Tagged:	5471
- NLP:	1410
- Text:	3891
- PDF Metadata:	67
- PDF Form Fields:	41
- Zonal Text:	30
- Zonal Barcode:	32

```
[Metadata identified for tagging]
   ⇨ NLP Extraction
    ▶ People
          • Maura McMahon
          • Christine Shaw
          • Charles Lamarche
          • Anna Friesel
          • Yong Hong
          • Chen
          • James Hsu
          • Dorian Cojocaru


DOCUMENT STATISTICS

Metadata Identified
    NLP   Text   PDF Form Fields    PDF Metadata    Zonal Text    Zonal
    Barcode
     8      0               0               0             0
     0

Metadata Tagged
    NLP   Text   PDF Form Fields    PDF Metadata    Zonal Text    Zonal
    Barcode
     8      0               0               0             0
     0
```

Service Status : Running

1 The **Statistics** panel shows a summary of all the documents processed so far (including previous Runs).



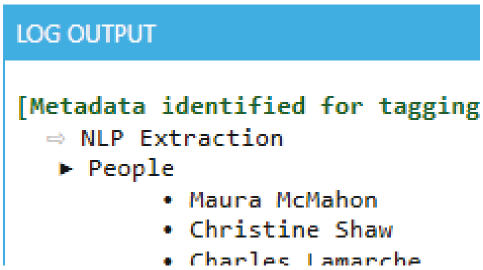
STATISTICS

Statistics for Entire Job

```
Total Processes:         8
 - Successful:           8
 - Partially successful: 1
 - Failed:               3

Total Metadata Tagged:   5
```

2 Sometimes, if there is an unexpected error or the computer restarts while a job is running, the statistics shown may not be accurate. Click on the **Re-calculate statistics** to refresh the values from the database.

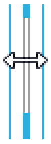
3 This shows the full log file. It is updated live as the job runs. See section 4.10.1.1 about how to analyse the log output



LOG OUTPUT

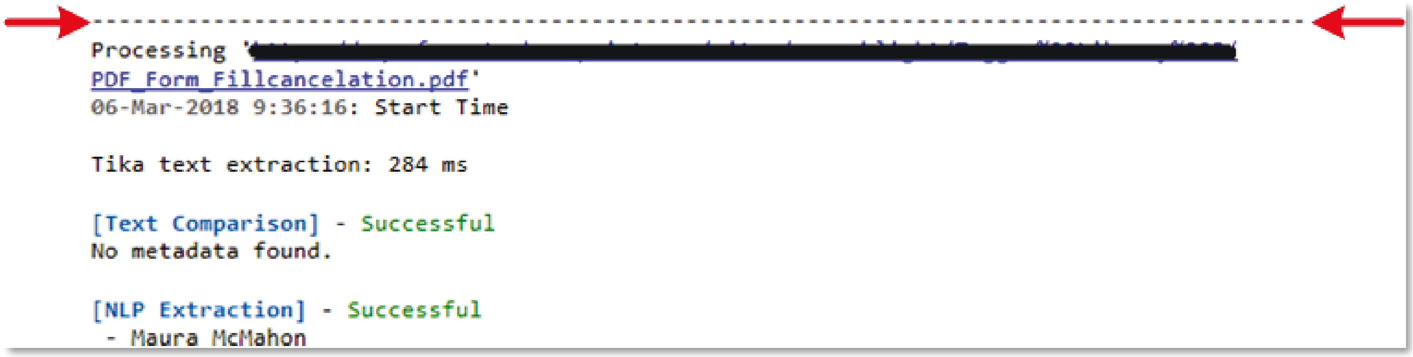
```
[Metadata identified for tagging
  ⇨ NLP Extraction
   ▸ People
           • Maura McMahon
           • Christine Shaw
           • Charles Lamarche
```

4 You can resize the **STATISTICS** and **LOG OUTPUT** panel



4.10.1.1 Analysing the log output

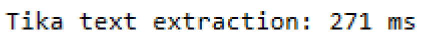
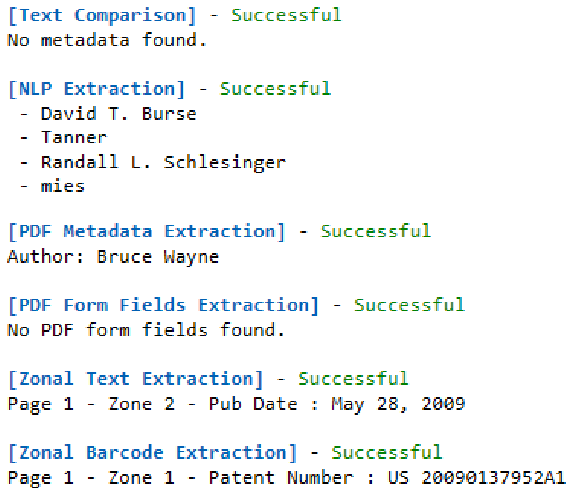
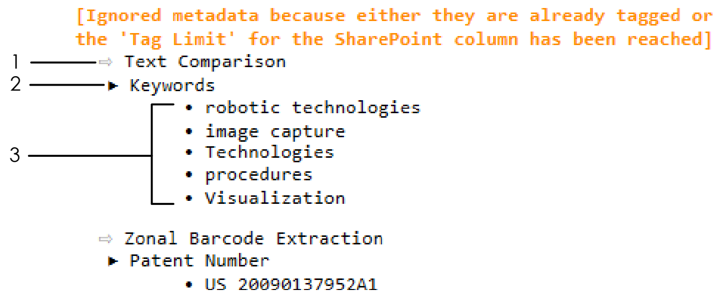
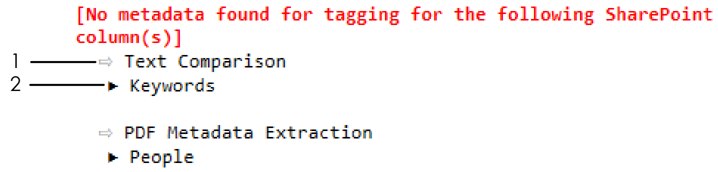
The log output for each processed document will be between dotted lines



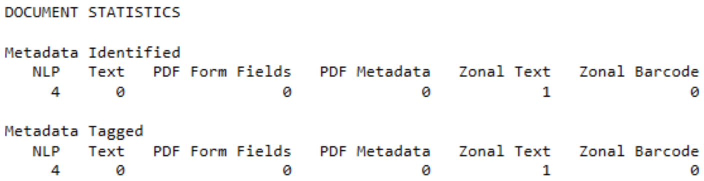
```
-------------------------------------------------------------------------------------------
Processing '[redacted]
PDF_Form_Fillcancelation.pdf'
06-Mar-2018 9:36:16: Start Time

Tika text extraction: 284 ms

[Text Comparison] - Successful
No metadata found.

[NLP Extraction] - Successful
 - Maura McMahon
```

Below is a list of outputs that (may) appear in the log output and what they mean.

Ouput	Description
`06-Mar-2018 9:36:16: Start Time`	The date and time the processing started for this file

Ouput	Description
 Tika text extraction: 271 ms	The time it took to extract and chunk the document's text
 [Text Comparison] - Successful No metadata found. [NLP Extraction] - Successful - David T. Burse - Tanner - Randall L. Schlesinger - mies [PDF Metadata Extraction] - Successful Author: Bruce Wayne [PDF Form Fields Extraction] - Successful No PDF form fields found. [Zonal Text Extraction] - Successful Page 1 - Zone 2 - Pub Date : May 28, 2009 [Zonal Barcode Extraction] - Successful Page 1 - Zone 1 - Patent Number : US 20090137952A1	The extraction results for each extraction task specified (for the job) and whether they were successful or not. The result(s) of the extraction are shown under each task. If a task did not return any results, it will still mark it as successful. It is only when there is an error when attempting to extract metadata from the document that it will mark it as being unsuccessful.
 [Ignored metadata because either they are already tagged or the 'Tag Limit' for the SharePoint column has been reached] ⇨ Text Comparison ► Keywords • robotic technologies • image capture • Technologies • procedures • Visualization ⇨ Zonal Barcode Extraction ► Patent Number • US 20090137952A1	You will see in the log file if any metadata extracted from any of the tasks are already present in SharePoint for this document 1. This shows the task that identified the metadata already in SharePoint 2. This is the SharePoint column that contain the metadata 3. This shows the list of metadata already in SharePoint You will also receive this if the Tag Limit for a SharePoint column has been reached. See section 5.5 for more information on tag limits.
 [No metadata found for tagging for the following SharePoint column(s)] ⇨ Text Comparison ► Keywords ⇨ PDF Metadata Extraction ► People	You will see this if no metadata is extracted for specific SharePoint columns after trying to extract using the defined task. 1. This shows the task that could not extract any metadata 2. The SharePoint column for which no metadata was extracted

<table>
<tr><th>Ouput</th><th>Description</th></tr>
<tr><td>



```
DOCUMENT STATISTICS

Metadata Identified
   NLP   Text   PDF Form Fields   PDF Metadata   Zonal Text   Zonal Barcode
     4      0                 0              0            1               0

Metadata Tagged
   NLP   Text   PDF Form Fields   PDF Metadata   Zonal Text   Zonal Barcode
     4      0                 0              0            1               0
```

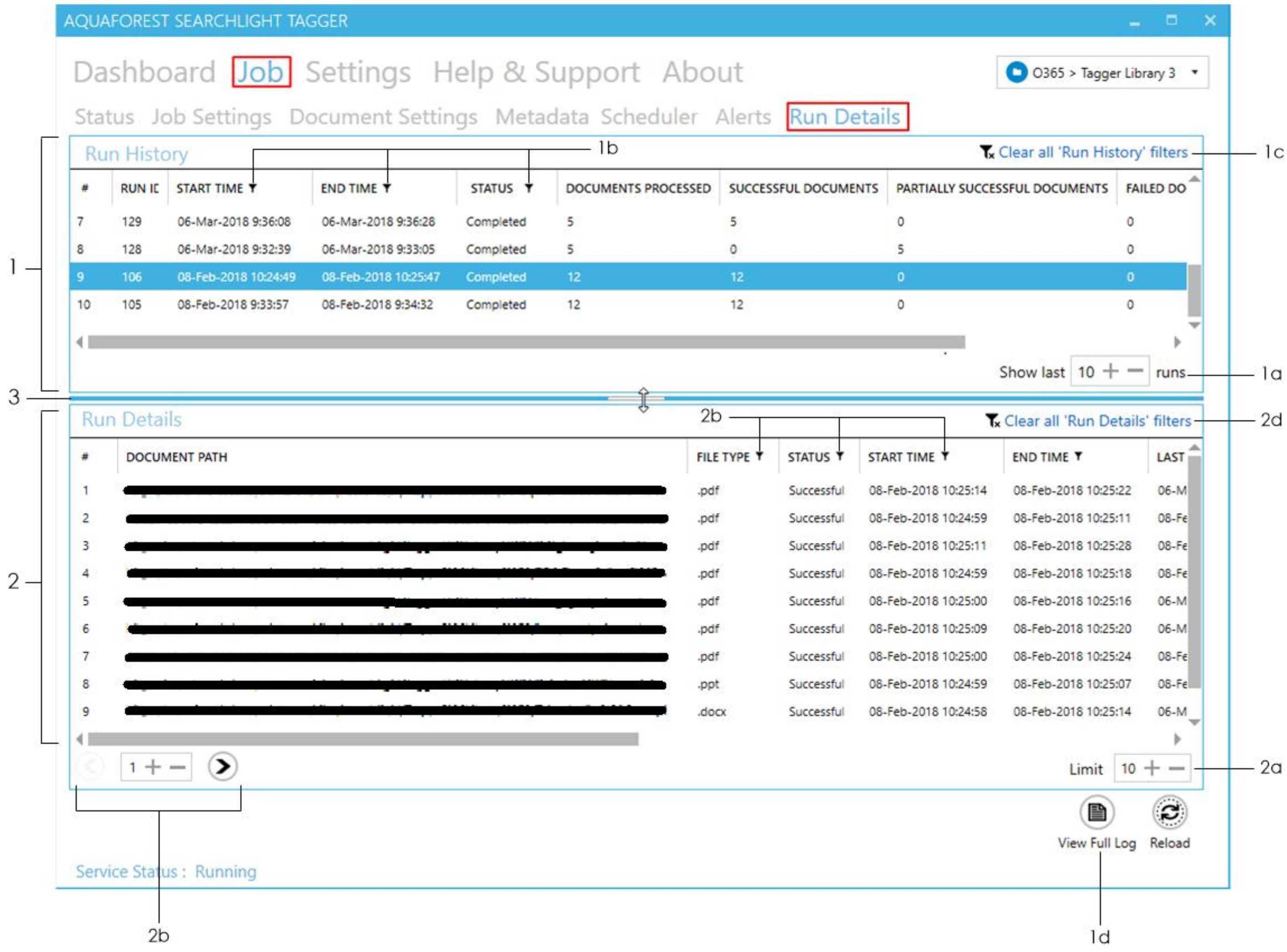
</td><td>This is a summary of the number of metadata that was identified (extracted) by each task and how many of those identified metadata were actually tagged.

If the number of metadata identified(extracted) is more than the number tagged, it could be because some of the metadata extracted were already tagged or the Tag Limit for some of the SharePoint columns was reached</td></tr>
<tr><td></td><td></td></tr>
</table>

For more detailed analysis of a job, go to Job > **Run Details** tab. See section 4.10.2 for more information.

4.10.2 Run Details

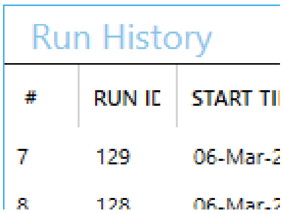
The Run Details tab enables detailed analysis of previous Job runs. To access it go to **Job** > **Run Details**.



#	RUN ID	START TIME	END TIME	STATUS	DOCUMENTS PROCESSED	SUCCESSFUL DOCUMENTS	PARTIALLY SUCCESSFUL DOCUMENTS	FAILED DO
7	129	06-Mar-2018 9:36:08	06-Mar-2018 9:36:28	Completed	5	5	0	0
8	128	06-Mar-2018 9:32:39	06-Mar-2018 9:33:05	Completed	5	0	5	0
9	106	08-Feb-2018 10:24:49	08-Feb-2018 10:25:47	Completed	12	12	0	0
10	105	08-Feb-2018 9:33:57	08-Feb-2018 9:34:32	Completed	12	12	0	0

#	DOCUMENT PATH	FILE TYPE	STATUS	START TIME	END TIME	LAST
1	[redacted]	.pdf	Successful	08-Feb-2018 10:25:14	08-Feb-2018 10:25:22	06-M
2	[redacted]	.pdf	Successful	08-Feb-2018 10:24:59	08-Feb-2018 10:25:11	08-Fe
3	[redacted]	.pdf	Successful	08-Feb-2018 10:25:11	08-Feb-2018 10:25:28	08-Fe
4	[redacted]	.pdf	Successful	08-Feb-2018 10:24:59	08-Feb-2018 10:25:18	08-Fe
5	[redacted]	.pdf	Successful	08-Feb-2018 10:25:00	08-Feb-2018 10:25:16	06-M
6	[redacted]	.pdf	Successful	08-Feb-2018 10:25:09	08-Feb-2018 10:25:20	06-M
7	[redacted]	.pdf	Successful	08-Feb-2018 10:25:00	08-Feb-2018 10:25:24	08-Fe
8	[redacted]	.ppt	Successful	08-Feb-2018 10:24:59	08-Feb-2018 10:25:07	08-Fe
9	[redacted]	.docx	Successful	08-Feb-2018 10:24:58	08-Feb-2018 10:25:14	06-M

1 The **Run History** panel shows all the runs so far.



#	RUN ID	START TI
7	129	06-Mar-2
8	128	06-Mar-2

a. You can choose how many runs to show. Runs are displayed in descending order of run date/time, that is, it will show the *last* 'x' runs.

Show last 10 + — runs

You need to click on the **Reload** button at the bottom of the window after updating this value.

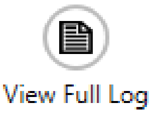
Reload

b. You can filter the Run History by any column that has the ▼ icon next to it. You can apply multiple filters by filtering each column one by one.

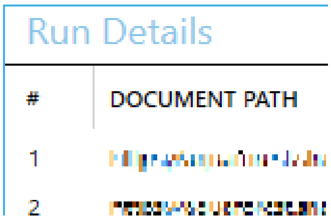
c. Use this to clear all filters currently applied.

Clear all 'Run History' filters

d. Click this to view the log file of the selected run.



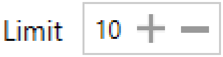
2 The **Run Details** panel shows all the documents that were processed for the selected run in the **Run History** panel.



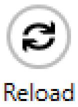
Run Details

#	DOCUMENT PATH
1	[redacted]
2	[redacted]

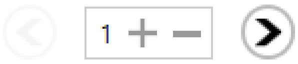
a. You can limit the number of documents to display per page.



You need to click on the **Reload** button at the bottom of the window after updating this value.

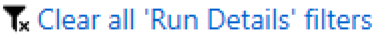


b. Display the next/previous 10 documents (since **Limit** is set to 10).

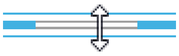


c. You can filter the Run Details by any column that has the ▼ icon next to it. You can apply multiple filters by filtering each column one by one.

d. Use this to clear all filters currently applied

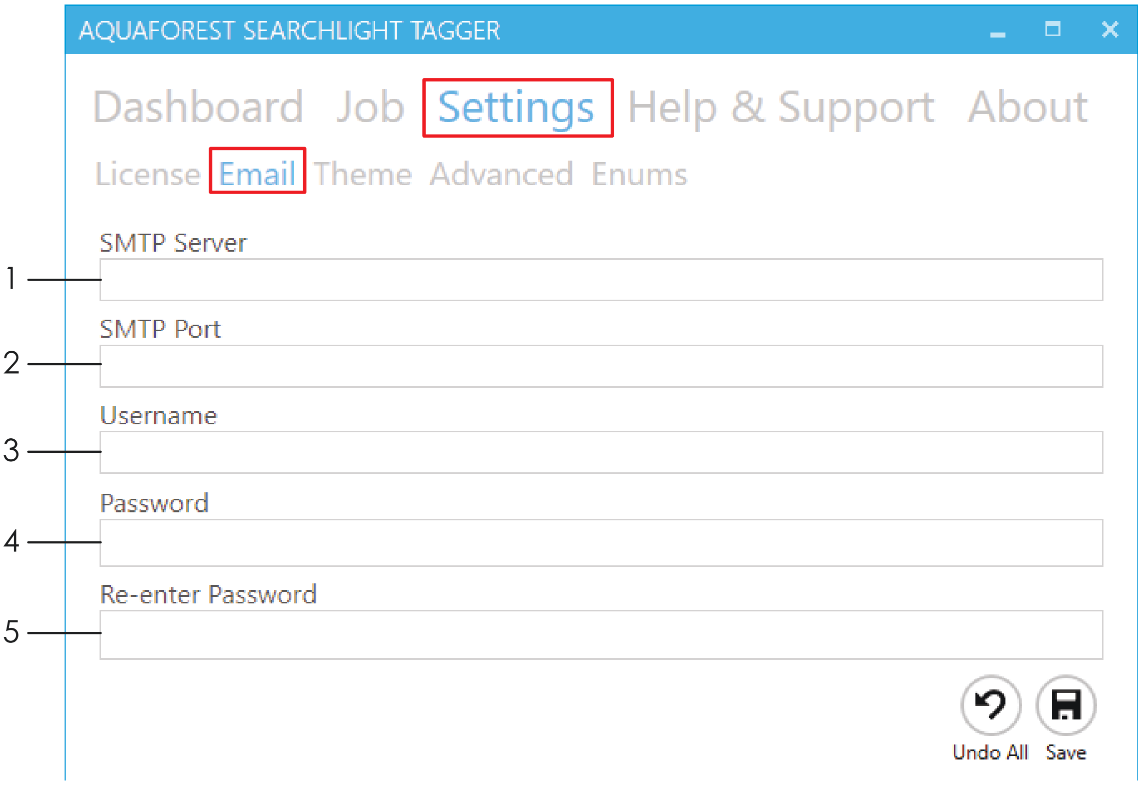


3 You can resize the **Run History** and **Run Details** panel

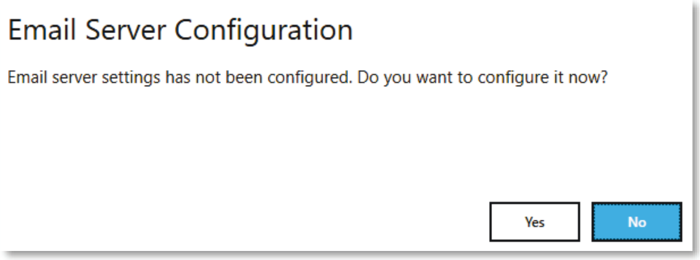


4.11 Email Settings

The Email settings tab allows email server information to be configured. This is used to send Alerts. To change these settings, go to **Settings** > **Email** tab.



Alternatively, if the email settings have not been set, you will be shown the following message and it you click on **Yes**, you will be shown a popup dialog box with the email settings shown above.



Email Server Configuration

Email server settings has not been configured. Do you want to configure it now?

1 The address of the server hosting the SMTP server

2 The SMTP Server port

3 Username for authentication by the server

4 Password for the username

4.12 Config file

The **Tagger.config** file contains advanced settings that should only be updated from guidance of the support team (support@aquaforest.com). The file is located in the following location: "{installation path}\config\Tagger.config".

If a setting in the config file is updated, the Tagger service must be restarted by going to **Settings > Advanced** and turn the service off and on again.

Some of the common settings available in the Tagger.config file are described below.

Setting	Description
debugLogging	Set this to true to help debug problematic errors. The debug log files will be saved to "{installation path}\live\log\{Job ID}\{JobID}[{RunID}]_debug.txt"
checkServiceEvery	This interval to periodically check the status of the Tagger service. If the status of a job is set to as running even though the service has stopped, it will be put into an error state. The default is to check the service every 60 minutes.
enumerationMaxParallelism	When enumerating documents from large SharePoint libraries, Aquaforest Searchlight Tagger partitions the retrieval so that the documents are retrieved in chunks. These chunks can be retrieved in parallel, which can significantly speed up enumeration. This setting is used to control the maximum number of chunks that can be retrieved at once. Note, however, that the maximum value will be limited to the maximum cores your license permits.
skipCheckedOutDocument	Set this to true to skip checked-out documents from being processed.
retainApprovalStatus	When documents are processed in a SharePoint library which requires Content Approval, it will set them to 'Pending' after processing. Set this setting to "true" to retain the original Approval Status after the documents have been processed.
downloadAndUploadRetries sharePointRequestRetries	Occasionally, there might be some intermittent network problems or unusual extreme load on the SharePoint server which can cause problems when processing SharePoint document libraries. To cope with this, retry mechanisms have been implemented for different scenarios that will retry performing a particular task in the event of such problems (e.g. timeouts). There are 2 SharePoint retry settings available: • downloadAndUploadRetries - used when downloading and uploading documents fail • sharePointRequestRetries - used when executing SharePoint queries fail The number of retries and the amount of time to wait between retries can be controlled through the respective config settings. The value needs to be entered in the format "x,y", where x is the number of retries and y is the time (in milliseconds) to wait before the first retry). For subsequent retries, the time to wait will be twice the previous wait time.

Setting	Description
databaseRetries	Sometimes, if a job is set to process using multiple cores, Tagger may encounter problems when it tries to update the database due to it being ‘locked’ because of concurrent updates. To overcome this problem, a retry mechanism has been implemented that will retry updating the database if it fails the first time. The number of retries and the amount of time to wait between retries can be controlled through this setting. The value needs to be entered in the format “x,y”, where x is the number of retries and y is the amount of time in milliseconds to wait for each retry.
formsAuthCookieRefreshInterval	The amount of time before refreshing forms based authentication cookies. The default is current set to 900,000 milliseconds (15 minutes).

5 Tips and FAQ

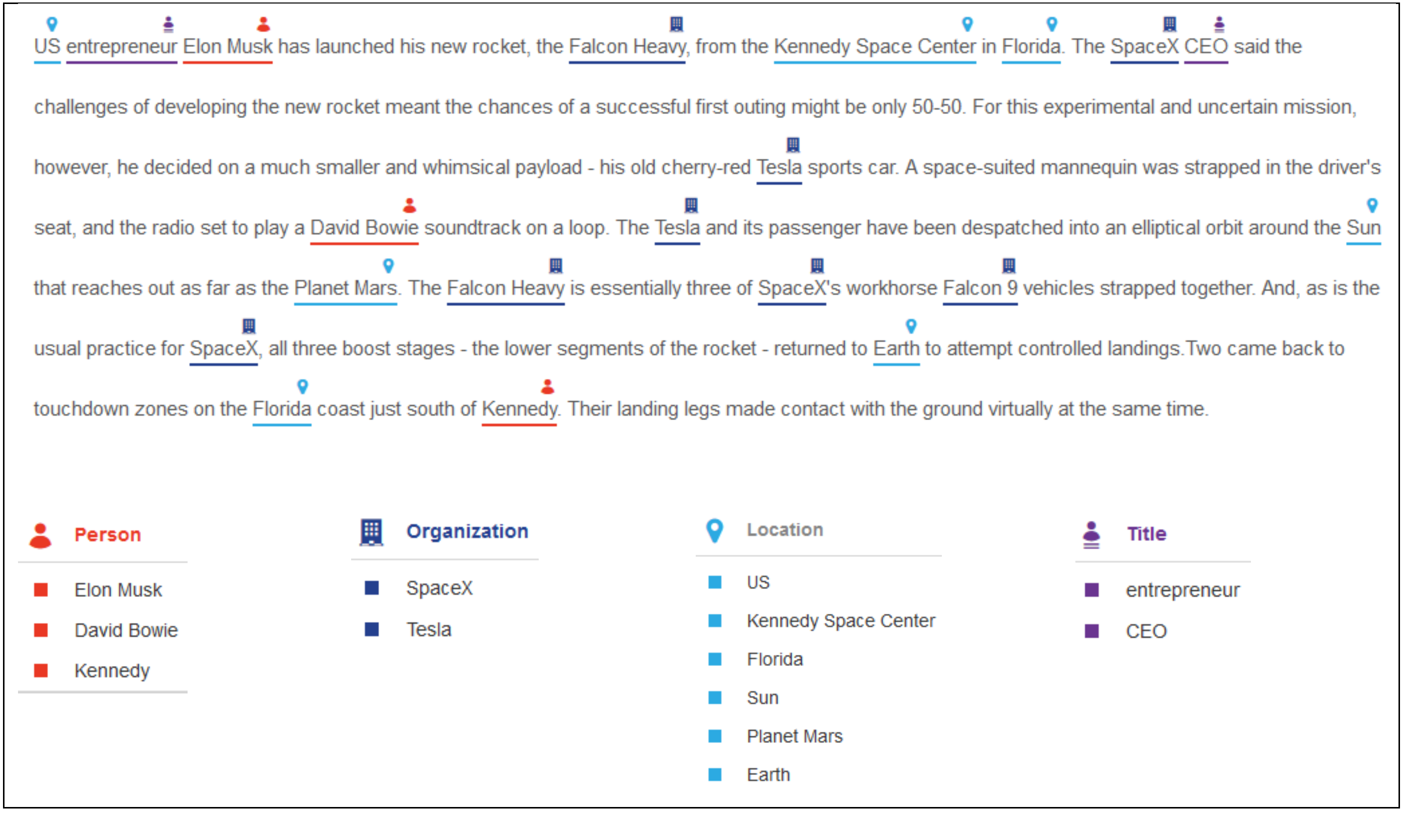
5.1 Entity Extraction (NLP)

Entity extraction is the process of automatically extracting named entities such as people, place, companies, etc. from unstructured contents in documents using Natural Language Processing (NLP).

Say, for example, we have the following text:

> US entrepreneur Elon Musk has launched his new rocket, the Falcon Heavy, from the Kennedy Space Center in Florida. The SpaceX CEO said the challenges of developing the new rocket meant the chances of a successful first outing might be only 50-50.
>
> For this experimental and uncertain mission, however, he decided on a much smaller and whimsical payload - his old cherry-red Tesla sports car. A space-suited mannequin was strapped in the driver's seat, and the radio set to play a David Bowie soundtrack on a loop. The Tesla and its passenger have been despatched into an elliptical orbit around the Sun that reaches out as far as the Planet Mars.
>
> The Falcon Heavy is essentially three of SpaceX's workhorse Falcon 9 vehicles strapped together. And, as is the usual practice for SpaceX, all three boost stages - the lower segments of the rocket - returned to Earth to attempt controlled landings. Two came back to touchdown zones on the Florida coast just south of Kennedy. Their landing legs made contact with the ground virtually at the same time.

This is the result of passing it to an NLP service:



US entrepreneur Elon Musk has launched his new rocket, the Falcon Heavy, from the Kennedy Space Center in Florida. The SpaceX CEO said the challenges of developing the new rocket meant the chances of a successful first outing might be only 50-50. For this experimental and uncertain mission, however, he decided on a much smaller and whimsical payload - his old cherry-red Tesla sports car. A space-suited mannequin was strapped in the driver's seat, and the radio set to play a David Bowie soundtrack on a loop. The Tesla and its passenger have been despatched into an elliptical orbit around the Sun that reaches out as far as the Planet Mars. The Falcon Heavy is essentially three of SpaceX's workhorse Falcon 9 vehicles strapped together. And, as is the usual practice for SpaceX, all three boost stages - the lower segments of the rocket - returned to Earth to attempt controlled landings.Two came back to touchdown zones on the Florida coast just south of Kennedy. Their landing legs made contact with the ground virtually at the same time.

Person	Organization	Location	Title
Elon Musk	SpaceX	US	entrepreneur
David Bowie	Tesla	Kennedy Space Center	CEO
Kennedy		Florida	
		Sun	
		Planet Mars	
		Earth	

The NLP service automatically identified Person, Organization, Location and Title from the text. If the text had other entity types, they would have been extracted too. Without NLP, the identification of these entities would have to have been done manually, which is not feasible for large number of documents in businesses.

The benefits of automated entity extraction for businesses are innumerable – from improving the finding of documents through faceted search (by categorising documents based on the entities) to unlocking valuable business related information that may otherwise be ‘hidden’.

5.1.1 Entity Extraction in Tagger

Aquaforest Searchlight Tagger is able to easily harness the power of automated entity extraction by using external third-party NLP service providers. To put it briefly, it is able to achieve this by first extracting the text from documents and then sending them over to the NLP service for processing. The results are then sent back to Tagger where they are processed further and eventually added to SharePoint as metadata. See section 1.2.3 for a diagrammatic representation of this.

In the current version, the following NLP services are supported:

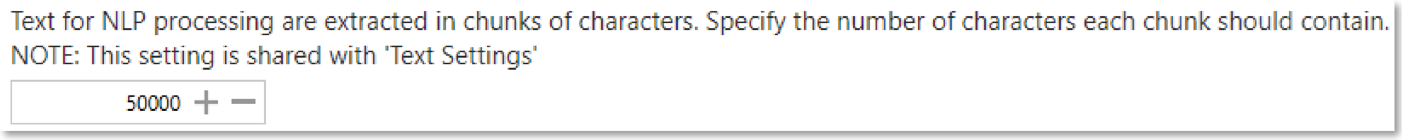
- Rosette
- Open Calais
- Microsoft Cognitive Services
- Google Natural Language

At the time of writing, all of the above NLP services offer free usage of their service. However, they come with certain restrictions as shown below.

NLP Service (free version)	Max API Calls	Text limit per call	
Rosette	10,000 calls per month 1,000 calls per day	600KB (50,000 characters)	more info
Open Calais	5,000 calls per day	100KB	
Microsoft Cognitive Services			
Google Natural Language	5,000 calls per month	1,000 characters	more info

5.1.1.1 Text Limit

Since, the free versions of each NLP service restricts the amount of text it can process at any one time, before sending a document's contents to the NLP service, Tagger split them in chunks of 50,000 characters. From our test, this seems to work for most NLP services currently supported. However, you can increase this value if you purchase their premium service. The following setting in Tagger under **Job** > **Metadata** > **NLP Settings** controls this:


Text for NLP processing are extracted in chunks of characters. Specify the number of characters each chunk should contain.
NOTE: This setting is shared with 'Text Settings'
50000 + −

5.1.1.2 API Calls

For every chunk that is sent to the NLP service, **1 API call** is consumed. You should schedule and limit the amount of documents processed to avoid going over the limit based on the selected NLP service.

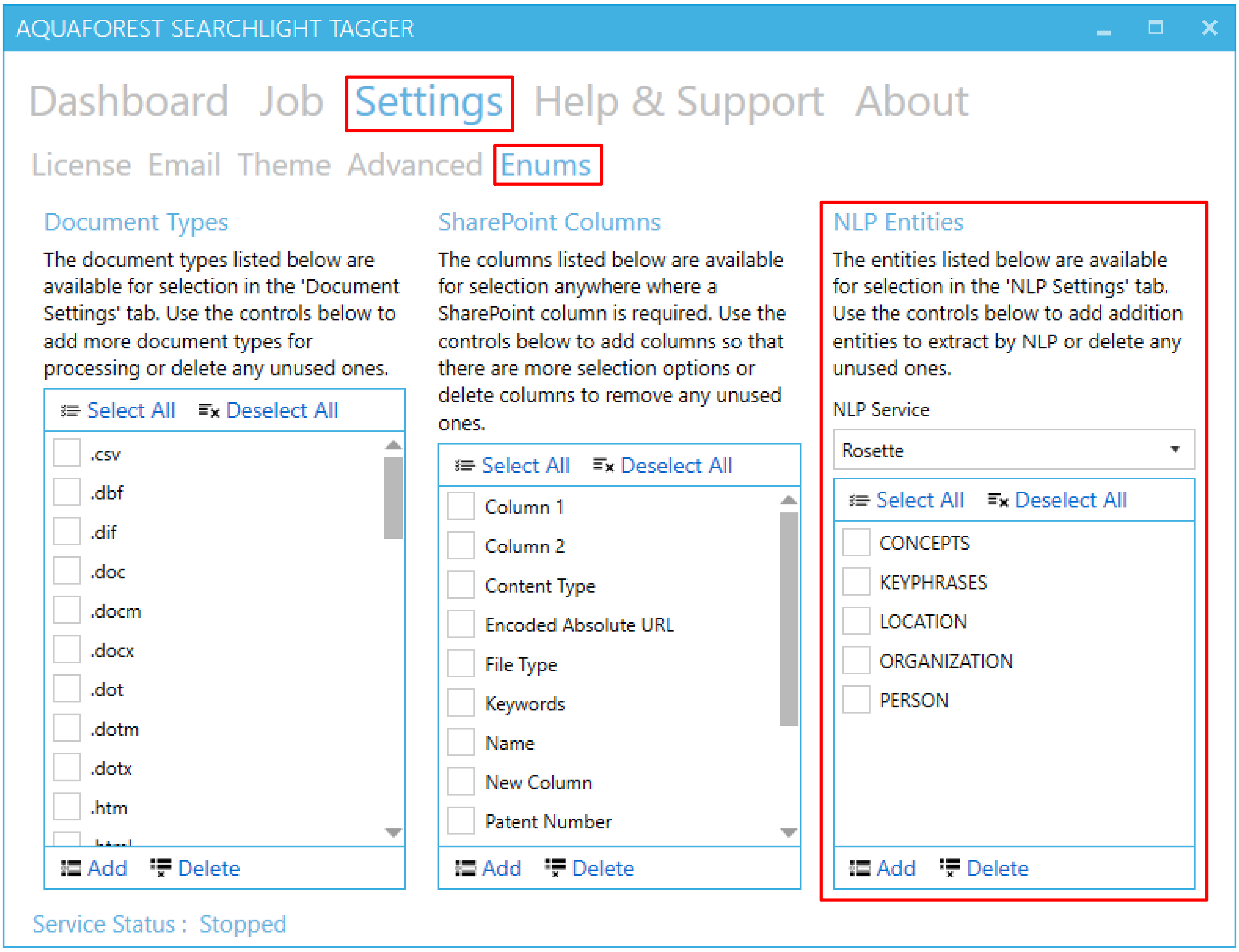
5.1.1.3 Entities

Each NLP service has its own entities that can be extracted. Tagger has the most common ones for each service.

NLP Service	Default Entities in Tagger	Additional entity types
Rosette	LOCATION ORGANIZATION PERSON CONCEPTS KEYPHRASES	more info

NLP Service	Default Entities in Tagger	Additional entity types
Open Calais	Country Company Person	additionalcontactdetails industry socialtags topic more info
Microsoft Cognitive Services	Keywords	
Google Natural Language	LOCATION ORGANIZATION PERSON	

To view the NLP entities currently defined in Tagger, go to **Settings** > **Enums** tab.



Document Types

The document types listed below are available for selection in the 'Document Settings' tab. Use the controls below to add more document types for processing or delete any unused ones.

SharePoint Columns

The columns listed below are available for selection anywhere where a SharePoint column is required. Use the controls below to add columns so that there are more selection options or delete columns to remove any unused ones.

NLP Entities

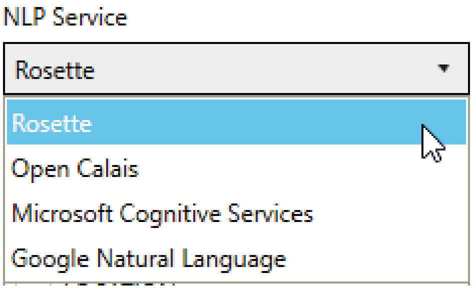
The entities listed below are available for selection in the 'NLP Settings' tab. Use the controls below to add addition entities to extract by NLP or delete any unused ones.

NLP Service: Rosette

Service Status : Stopped

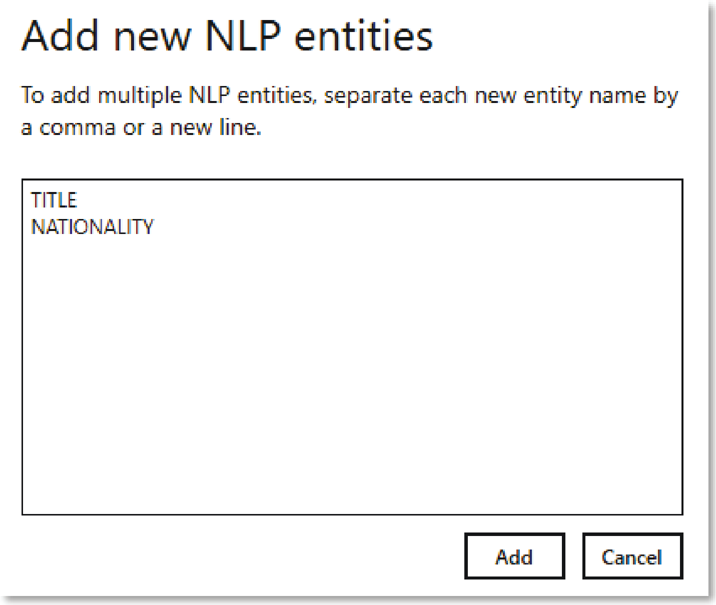
To add new entities

1 Select the NLP Service for which you want to add entities



2 Click on the Add button

3 A popup dialog will appear. Enter entity name(s).

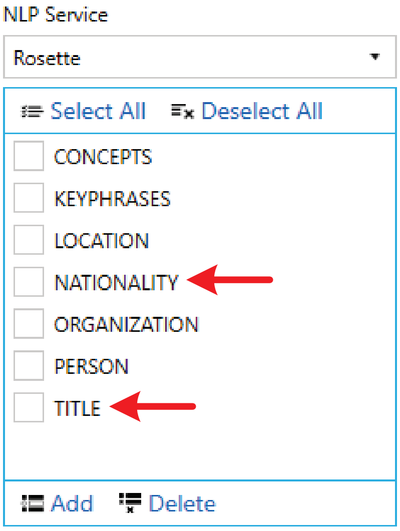


Add new NLP entities

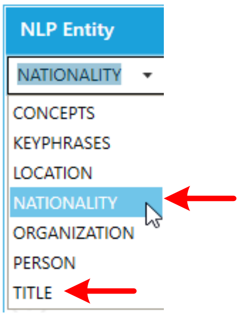
To add multiple NLP entities, separate each new entity name by a comma or a new line.

TITLE
NATIONALITY

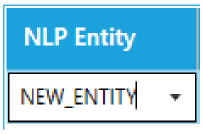
You can add multiple entities by separating each new entity by a comma or a new line. Click the **Add** button after specifying all the new entity names.



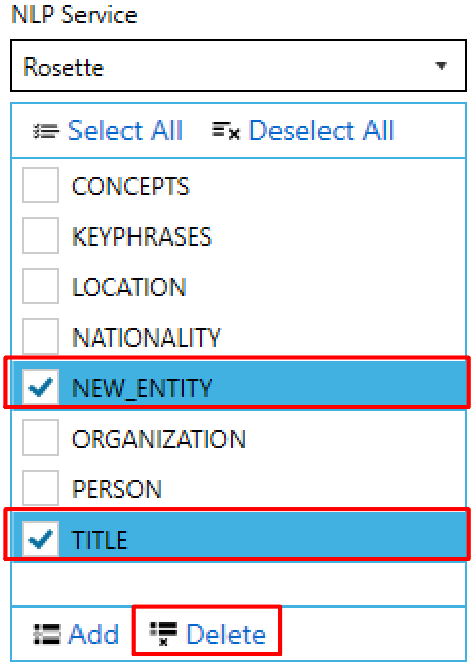
Now these entities will be available for selection under **Job** > **Metadata** > **NLP Settings**.



Another way to add a new entity is just type it in directly in the drop down menu.



You can also delete any unused entities. Select the entity(ies) you want to delete and click on the Delete button


NLP Service: Rosette

5.1.2 Generating API keys

In order to be able to extract entities from documents in Tagger, you need to create a free account with the NLP service you wish to use and generate an API key.

1 Go to **Job** > **Metadata** > **NLP Settings**

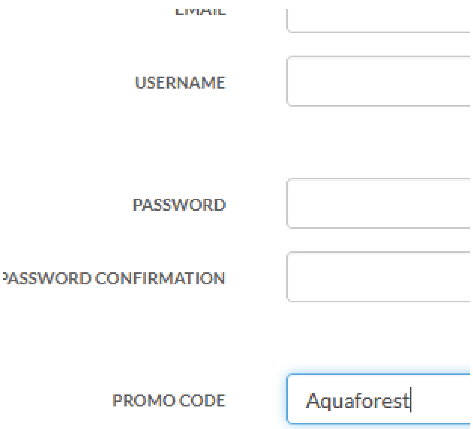
2 Select the NLP service you want to use

3 Click on Don't have a token? link next **Token/API Key to access NLP Service**

This will open the registration page for the selected NLP service in your default web browser. Complete the signup process:

a. **Rosette**:

https://developer.rosette.com/signup



EMAIL

USERNAME

PASSWORD

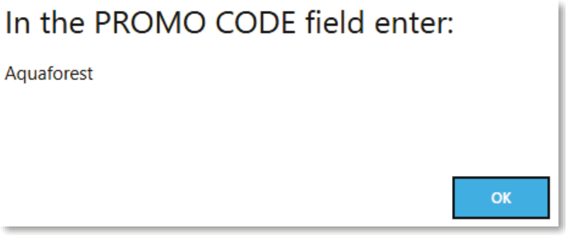
PASSWORD CONFIRMATION

PROMO CODE Aquaforest

IMPORTANT

When signing up for **Rosette**, make sure you add **Aquaforest** in the **PROMO CODE** field.

You will be notified via a message box to do it.



In the PROMO CODE field enter:

Aquaforest

b. **Open Calais**:

http://www.opencalais.com/opencalais-api/

c. **Microsoft Cognitive Service**

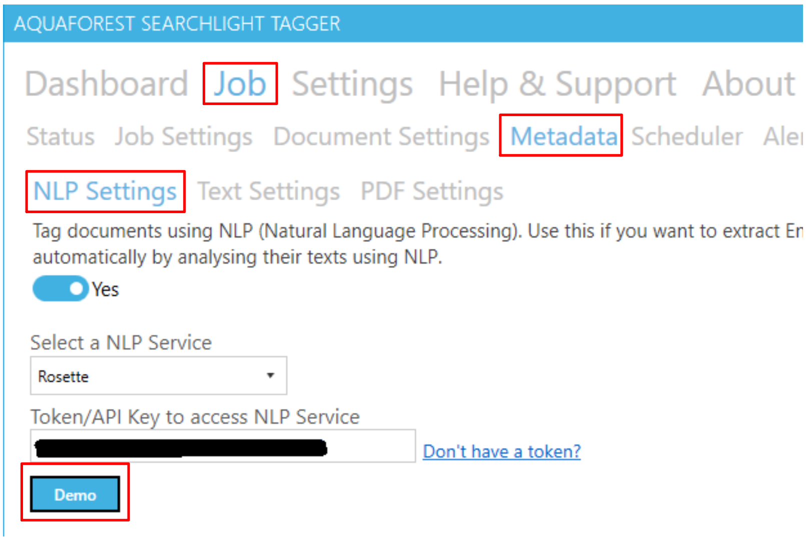
https://labs.cognitive.microsoft.com/en-us/project-entity-linking

d. **Google Natural Language**: https://console.cloud.google.com/freetrial

4 Once you receive the API key, enter it in the **Token/API Key to access NLP Service** textbox in Tagger.

5.1.3 Entity Extraction Demo

To quickly test if the API key is valid and working, click on the **Demo** button under **Job** > **Metadata** > **NLP Settings**.

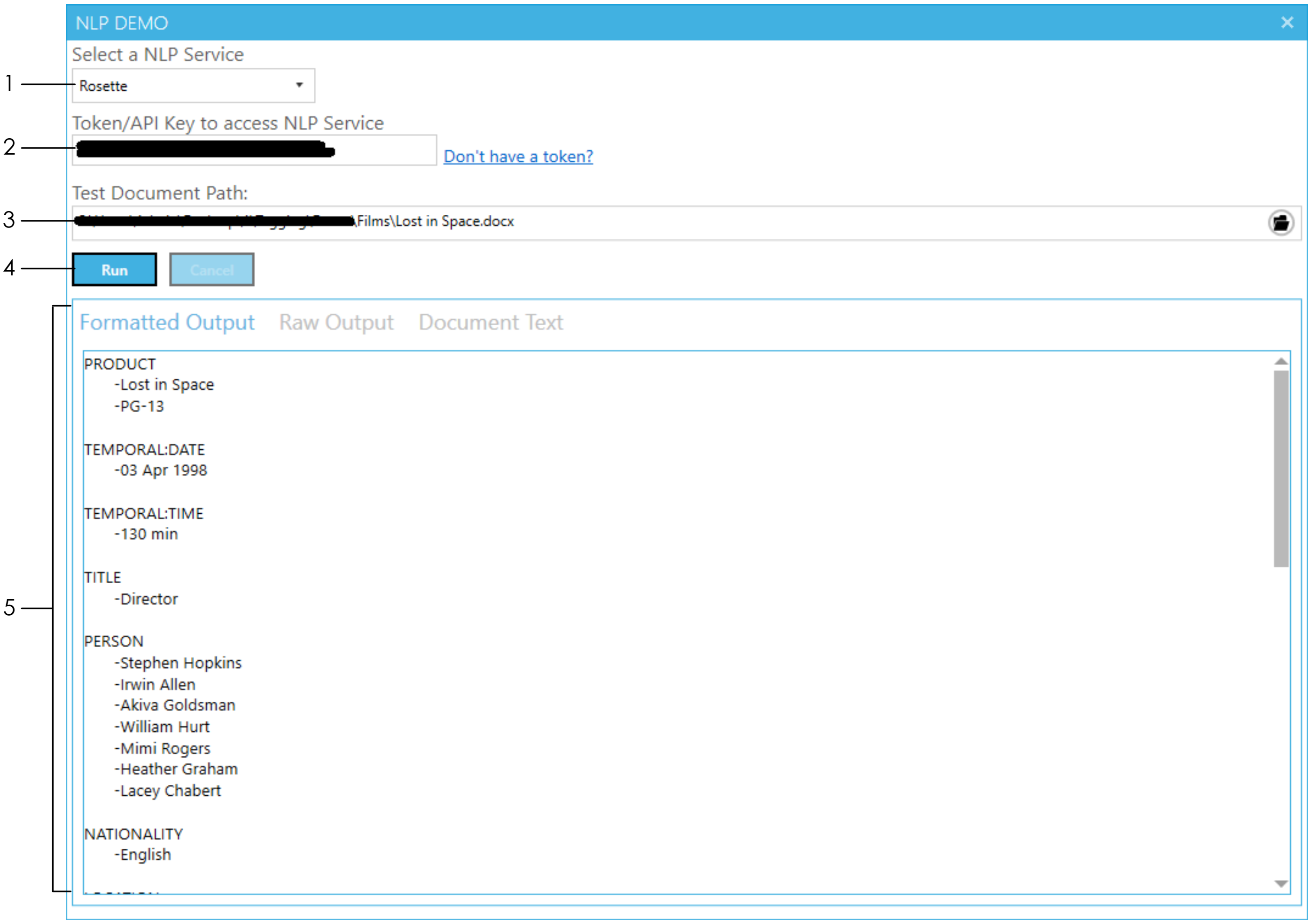


Tag documents using NLP (Natural Language Processing). Use this if you want to extract En automatically by analysing their texts using NLP.

● Yes

Select a NLP Service: Rosette

Token/API Key to access NLP Service: [redacted]



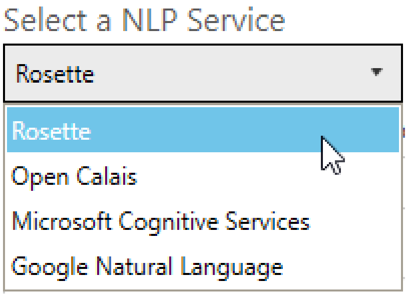
Select a NLP Service: Rosette

Token/API Key to access NLP Service: [redacted]

Test Document Path: [redacted]\Films\Lost in Space.docx

```
PRODUCT
      -Lost in Space
      -PG-13

TEMPORAL:DATE
      -03 Apr 1998

TEMPORAL:TIME
      -130 min

TITLE
      -Director

PERSON
      -Stephen Hopkins
      -Irwin Allen
      -Akiva Goldsman
      -William Hurt
      -Mimi Rogers
      -Heather Graham
      -Lacey Chabert

NATIONALITY
      -English
```

1 Select the NLP service you want to demo.



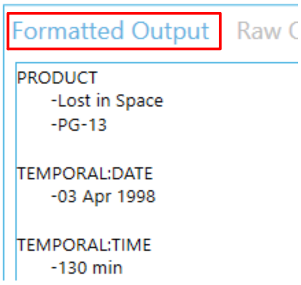
2 Enter the API key to access the selected NLP service. If you do not have an API key, generate one.

3 Select a sample file to use for the demo.

4 Click the **Run** button and wait for the NLP service to return the results.

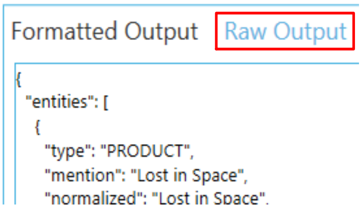
5

- The **Formatted Output** tab shows the extracted entities after Tagger has formatted them.



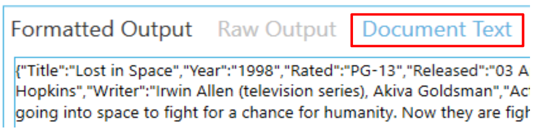
```
PRODUCT
    -Lost in Space
    -PG-13

TEMPORAL:DATE
    -03 Apr 1998

TEMPORAL:TIME
    -130 min
```

- To view the raw output as returned by the NLP service, click on **Raw Output** tab



```
{
 "entities": [
  {
   "type": "PRODUCT",
   "mention": "Lost in Space",
   "normalized": "Lost in Space",
```

- To view the text (from the document) that was sent to the NLP service, click on the **Document Text** tab



```
{"Title":"Lost in Space","Year":"1998","Rated":"PG-13","Released":"03 A
Hopkins","Writer":"Irwin Allen (television series), Akiva Goldsman","Ac
going into space to fight for a chance for humanity. Now they are figh
```

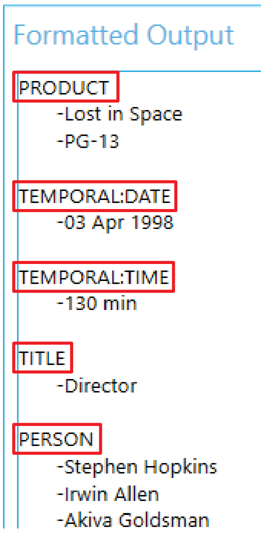
When using the demo, ***all*** entities supported by the NLP service are retrieved. This can be useful if you want to extract entities that are not part of the default ones provided and do not know the names of the other entities.

To view all the entities extracted from the document go to the **Formatted Output** tab.

The names of the entities are shown in red in the image. To add them:

1. Make a note of the ones you want to add
2. Close the Demo window
3. Go to **Settings** > **Enum** tab
4. Add them by following the instructions here

NOTE: Running the demo will also use up your API calls. So, be careful not to demo too many times and make sure to limit the number of chunks that is processed if you are testing large documents because they will be split into chunks and each chunk will consume one API call.



Formatted Output

PRODUCT
-Lost in Space
-PG-13

TEMPORAL:DATE
-03 Apr 1998

TEMPORAL:TIME
-130 min

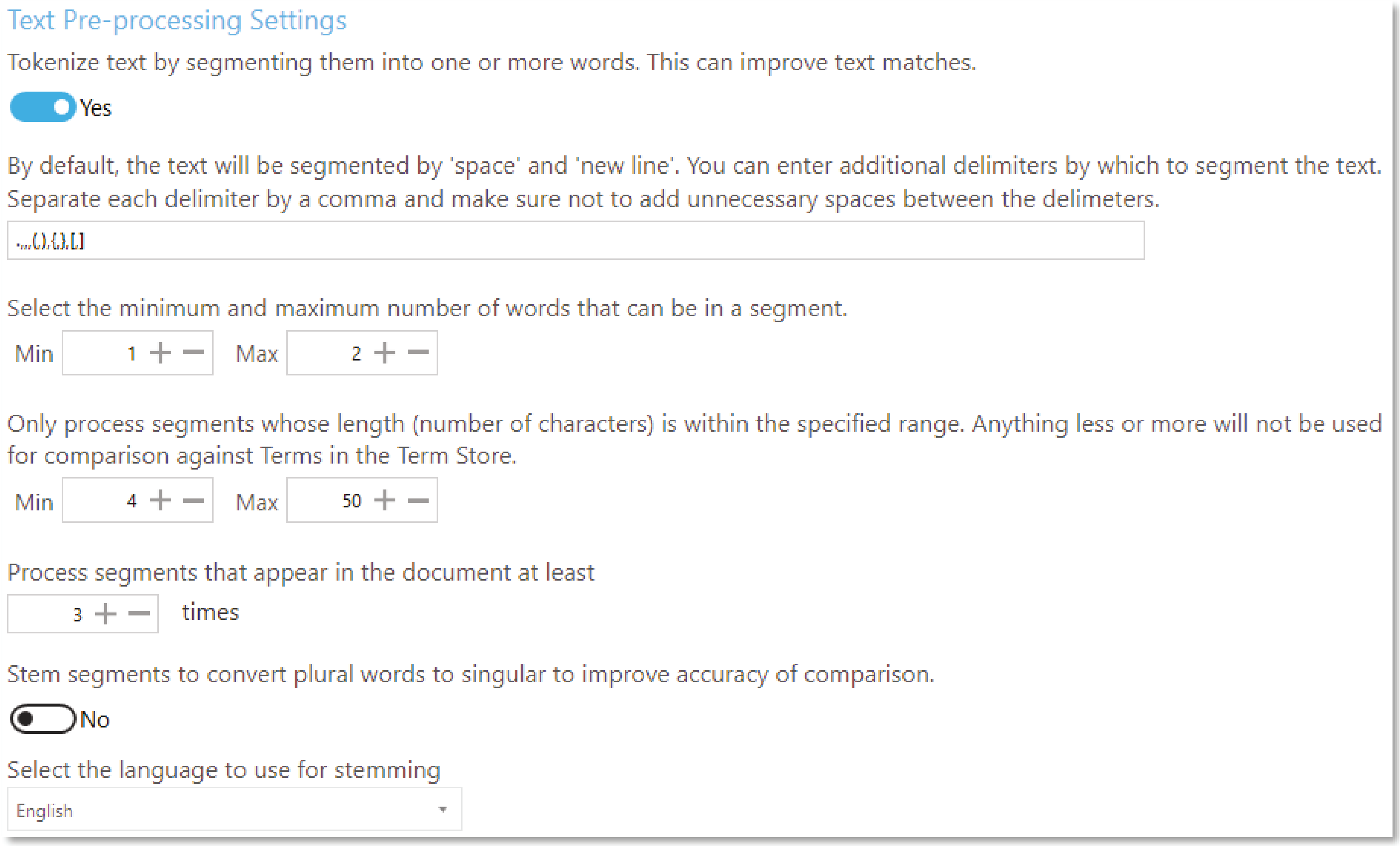
TITLE
-Director

PERSON
-Stephen Hopkins
-Irwin Allen
-Akiva Goldsman

5.2 Tokenization

Tokenization is the process of breaking text into individual words, phrases, symbols, etc. called tokens (or segments).

In Tagger, tokenization can be used for Taxonomy Matching and it is controlled by **Text Pre-processing Settings** under the **Job** > **Metadata** > **Taxonomy Matching Settings** tab.



Text Pre-processing Settings

Tokenize text by segmenting them into one or more words. This can improve text matches.

Yes

By default, the text will be segmented by 'space' and 'new line'. You can enter additional delimiters by which to segment the text. Separate each delimiter by a comma and make sure not to add unnecessary spaces between the delimeters.

',.,,(,),{,},[,]

Select the minimum and maximum number of words that can be in a segment.

Min 1 + − Max 2 + −

Only process segments whose length (number of characters) is within the specified range. Anything less or more will not be used for comparison against Terms in the Term Store.

Min 4 + − Max 50 + −

Process segments that appear in the document at least

3 + − times

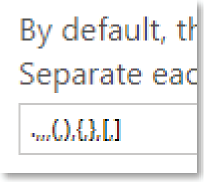
Stem segments to convert plural words to singular to improve accuracy of comparison.

No

Select the language to use for stemming

English

By default, Tagger will tokenize text using the *space* and *new line* characters but you can specify additional delimiters to use to tokenize text. When you create a new job, Tagger will have default additional delimiters as shown below.



By default, th
Separate eac

',.,,(,),{,},[,]

The delimiters (shown below in green) must be separated by a comma:

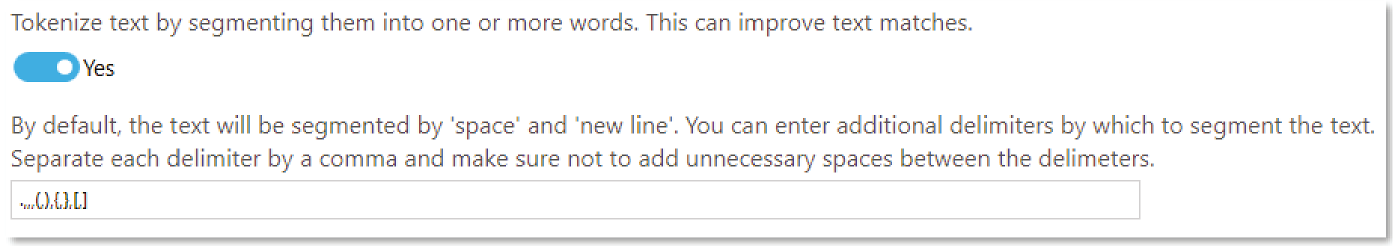
.,,,(,),{,},[,]

You can add or remove delimiters from the default values. Just make sure to avoid having unnecessary spaces between the delimiters.

Let us look at an example of how tokenization works in Tagger. Say we had the following text (adapted from *"The Everlasting Story of Nory"* by *"Nicholson Baker"*):

Nory was an ice cream vendor because her mother was an ice cream vendor, and Nory's mother was an ice cream vendor because her father was an ice cream vendor, and her father was an ice cream vendor because his mother was an ice cream vendor, or had been.

Based on the following Tagger settings,



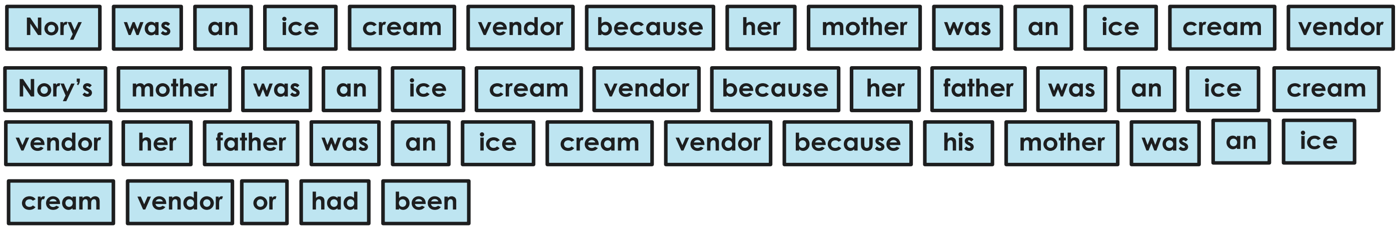
Tokenize text by segmenting them into one or more words. This can improve text matches.

Yes

By default, the text will be segmented by 'space' and 'new line'. You can enter additional delimiters by which to segment the text. Separate each delimiter by a comma and make sure not to add unnecessary spaces between the delimeters.

',.,(),{},[]

Tagger will tokenize the sentence as follows.

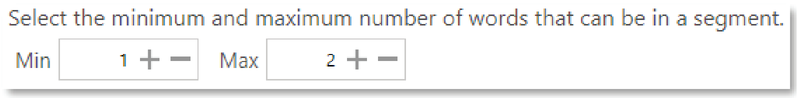


Nory was an ice cream vendor because her mother was an ice cream vendor
Nory's mother was an ice cream vendor because her father was an ice cream
vendor her father was an ice cream vendor because his mother was an ice
cream vendor or had been

Each 'box' is a token/segment. For this particular sentence, the tokens were generated after splitting it with *space*, *comma* and *full stop* delimiters because these are the only delimiters present in the sentence. Note how the delimiters are not part of the tokens.

With only these two Tagger settings, each of the tokens above will be compared to Terms in the SharePoint Term Store. However, this is not very efficient since there are quite a few duplicate tokens. Moreover, if the Term Set(s) being compared had the word "**ice cream**", the above sentence would not return a match because "**ice**" and "**cream**" are two separate tokens.

To deal with this Tagger has the following setting, which allows combining tokens together.



Select the minimum and maximum number of words that can be in a segment.

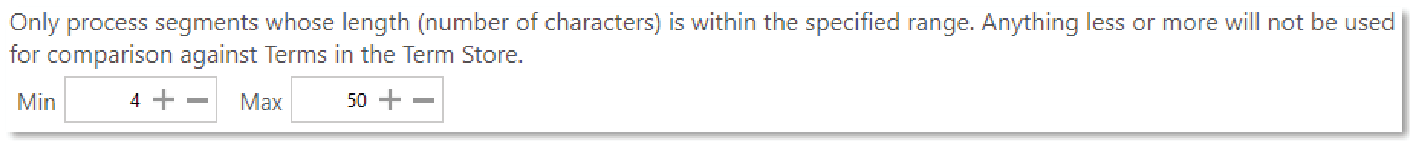
Min 1 + −   Max 2 + −

Using the above settings, Tagger will combine up to two tokens together resulting in the following:

Nory	x1	cream vendor	x6	vendor and	x2	because his	x1
Nory was	x1	vendor	x6	and	x2	his	x1
was	x6	vendor because	x3	and Nory’s	x1	his mother	x1
was an	x6	because	x3	Nory’s	x1	vendor or	x1
an	x6	because her	x2	Nory’s mother	x1	or	x1
an ice	x6	her	x3	her father	x2	or had	x1
ice	x6	her mother	x1	father	x2	had	x1
→ ice cream	x6	mother	x3	father was	x2	had been	x1
cream	x6	mother was	x3	and her	x1	been	x1

You will also notice that duplicate tokens have been grouped together to avoid comparing the same tokens multiple times.

The comparison process can be further optimized by excluding tokens of certain lengths to improve efficiency and effectiveness. This can be useful to remove common less pertinent words (or stop words) such as “a”, “or”, “to”, etc. The following settings control this:



Only process segments whose length (number of characters) is within the specified range. Anything less or more will not be used for comparison against Terms in the Term Store.

Min 4 + − Max 50 + −

Using the above settings, only tokens whose length is between 4 and 50 characters will be used for comparison. Consequently, the tokens shown in red below will be excluded because they are less than four characters.

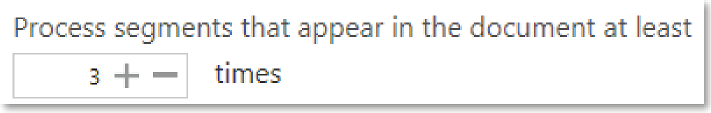
Nory	cream vendor	vendor and	because his
Nory was	vendor	🟥 and	🟥 his
🟥 was	vendor because	and Nory’s	his mother
was an	because	Nory’s	vendor or
🟥 an	because her	Nory’s mother	🟥 or
an ice	🟥 her	her father	or had
🟥 ice	her mother	father	🟥 had
ice cream	mother	father was	had been
cream	mother was	and her	been

In order to improve the accuracy and validity of terms tagged, we can tell Tagger to compare only those tokens that appear at least a minimum number of times.

Below are the remaining tokens and the frequency of their appearance in the sentence.

Nory	x1	cream vendor	x6	vendor and	x2	because his	x1
Nory was	x1	vendor	x6	and Nory's	x1	his mother	x1
was an	x6	vendor because	x3	Nory's	x1	vendor or	x1
an ice	x6	because	x3	Nory's mother	x1	or had	x1
ice cream	x6	because her	x2	her father	x2	had been	x1
cream	x6	her mother	x1	father	x2	been	x1
		mother	x3	father was	x2		
		mother was	x3	and her	x1		

Using the following setting, we can tell Tagger to only compare tokens that appear at least 3 times (in the document).



Process segments that appear in the document at least

3 + — times

Nory	x1	cream vendor	x6	vendor and	x2	because his	x1
Nory was	x1	vendor	x6	and Nory's	x1	his mother	x1
was an	x6	vendor because	x3	Nory's	x1	vendor or	x1
an ice	x6	because	x3	Nory's mother	x1	or had	x1
ice cream	x6	because her	x2	her father	x2	had been	x1
cream	x6	her mother	x1	father	x2	been	x1
		mother	x3	father was	x2		
		mother was	x3	and her	x1		

Consequently, only the following tokens will be compared against Terms in the Term Store:

was an	x6	vendor	x6
an ice	x6	vendor because	x3
ice cream	x6	because	x3
cream	x6	mother	x3
cream vendor	x6	mother was	x3

Using all the settings described above, Tagger can efficiently and accurately match text from documents to Terms in Term Store.

5.2.1 Stemming

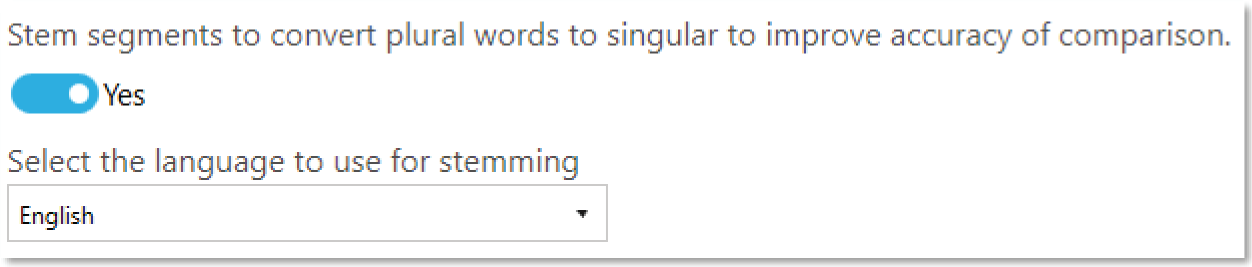
Stemming is the process of reducing words to their root form. Most languages have inflected version of words to express different grammatical categories such as number, tense, gender, mood, etc.

Example:

Root form	Inflected form(s)
Child	Children
Play	Playing Played
Engineer	Engineers Engineered Engineering

If the SharePoint Term Store has the root form of a word as a Term (e.g. Engineer), and a document has the inflected form of the word (e.g. Engineers), it will not match and therefore will not be tagged. Using stemming, Tagger will attempt to convert the inflected form in the document to its root form, which will match, thus improving comparison accuracy.

To use stemming in Tagger, enable it and set the language to use for stemming based on the language of the documents being processed because different languages have different stemming rules.



Stem segments to convert plural words to singular to improve accuracy of comparison.

● Yes

Select the language to use for stemming

English ▾

5.3 Patterns (Regular Expressions)

In Tagger, there a several places where you can specify patterns or regular expressions to constrain metadata that is extracted or tagged. Regular expressions enable you to apply formatting rules, check lengths, etc. to text to make sure they match a specific pattern. In essence, it validates the metadata before they are extracted from the document or tagged in SharePoint.

Here are some basic examples

Regular Expression	Example matches	Description
abc$	abc 123abc	Any text ending with abc
^abc	abc abc123	Any text that starts with abc
^[0-9]{5}$	11111 12345 99999	Any 5 digit numbers
\d{1,4}	1 24 445 3333	Any number that is 1 to 4 digits
[A-Za-z]{4}-\d{4}	ABCD-1234 GYDL-8450	4 letters followed by a dash, then 4 numbers
[A-Za-z]{4}(-| |_)\d{4}	ABCD-1234 ABCD_1234 ABCD 1234	4 letters followed either by a dash, underscore or space, then 4 numbers. It will not match the following: • ABCD>1234 • ABCD+1234
[A-Za-z]{4}[\W_]\d{4}	ABCD-1234 ABCD_1234 ABCD 1234 ABCD+1234 ABCD#1234	4 letters followed by any non-word separator, then 4 numbers

Below are a few useful resources to get you started with regular expressions:

- https://docs.microsoft.com/en-us/dotnet/standard/base-types/regular-expression-language-quick-reference
- https://msdn.microsoft.com/en-us/library/ms972966.aspx
- https://msdn.microsoft.com/en-us/library/ff650303.aspx

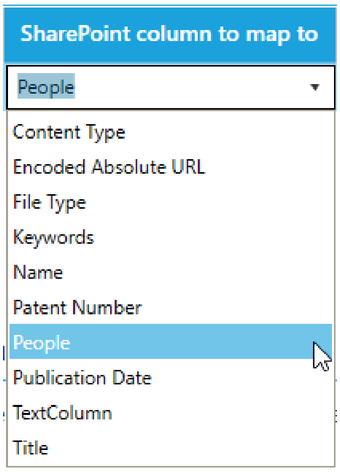
Some useful regular expressions taken from the resources above:

<table>
<tr><th>Field</th><th>Regular Expression</th><th>Example matches</th><th>Description</th></tr>
<tr><td>Social Security Number</td><td>^\d{3}-\d{2}-\d{4}$</td><td>111-11-1111</td><td>Validates the format, type, and length of the supplied input field. The input must consist of 3 numeric characters followed by a dash, then 2 numeric characters followed by a dash, and then 4 numeric characters.</td></tr>
<tr><td>Phone Number</td><td>^[01]?[- .]?(\([2-9]\d{2}\)|[2-9]\d{2})[- .]?\d{3}[- .]?\d{4}$</td><td>(425) 555-0123
425-555-0123
425 555 0123
1-425-555-0123</td><td>Validates a U.S. phone number. It must consist of 3 numeric characters, optionally enclosed in parentheses, followed by a set of 3 numeric characters and then a set of 4 numeric characters.</td></tr>
<tr><td>E-mail</td><td>^(?("")("".+?""@)|(([0-9a-zA-Z]((\.(?!\.))|[-!#\$%&'*\+/=\?\^`\{\}\|~\w])*)(?<=[0-9a-zA-Z])@))(?(\[)(\[(\d{1,3}\.){3}\d{1,3}\])|(([0-9a-zA-Z][-\w]*[0-9a-zA-Z]\.)+[a-zA-Z]{2,6}))$</td><td>someone@example.com</td><td>Validates an e-mail address.</td></tr>
<tr><td>ZIP Code</td><td>^(\d{5}-\d{4}|\d{5}|\d{9})$|^([a-zA-Z]\d[a-zA-Z] \d[a-zA-Z]\d)$</td><td>12345</td><td>Validates a U.S. ZIP Code. The code must consist of 5 or 9 numeric characters.</td></tr>
<tr><td>Currency (non-negative)</td><td>^\d+(\.\d\d)?$</td><td>1.00</td><td>Validates a positive currency amount. If there is a decimal point, it requires 2 numeric characters after the decimal point. For example, 3.00 is valid but 3.1 is not.</td></tr>
<tr><td>Currency (positive or negative)</td><td>^(-)?\d+(\.\d\d)?$</td><td>1.20</td><td>Validates for a positive or negative currency amount. If there is a decimal point, it requires 2 numeric characters after the decimal point.</td></tr>
</table>

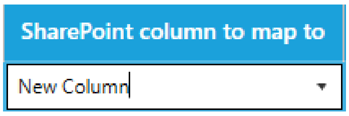
5.4 SharePoint Columns

Anywhere in Tagger where you are required to enter a SharePoint column, you will be provided with a drop-down menu.

For instance:

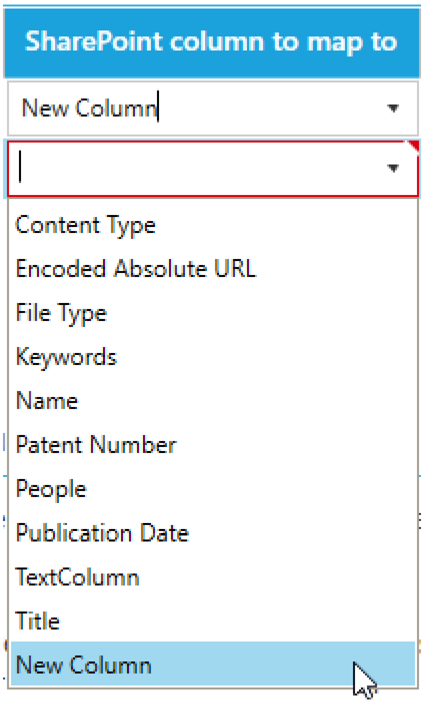


If you want to specify a column that is not present in the drop-down menu, you can type it in the column name.

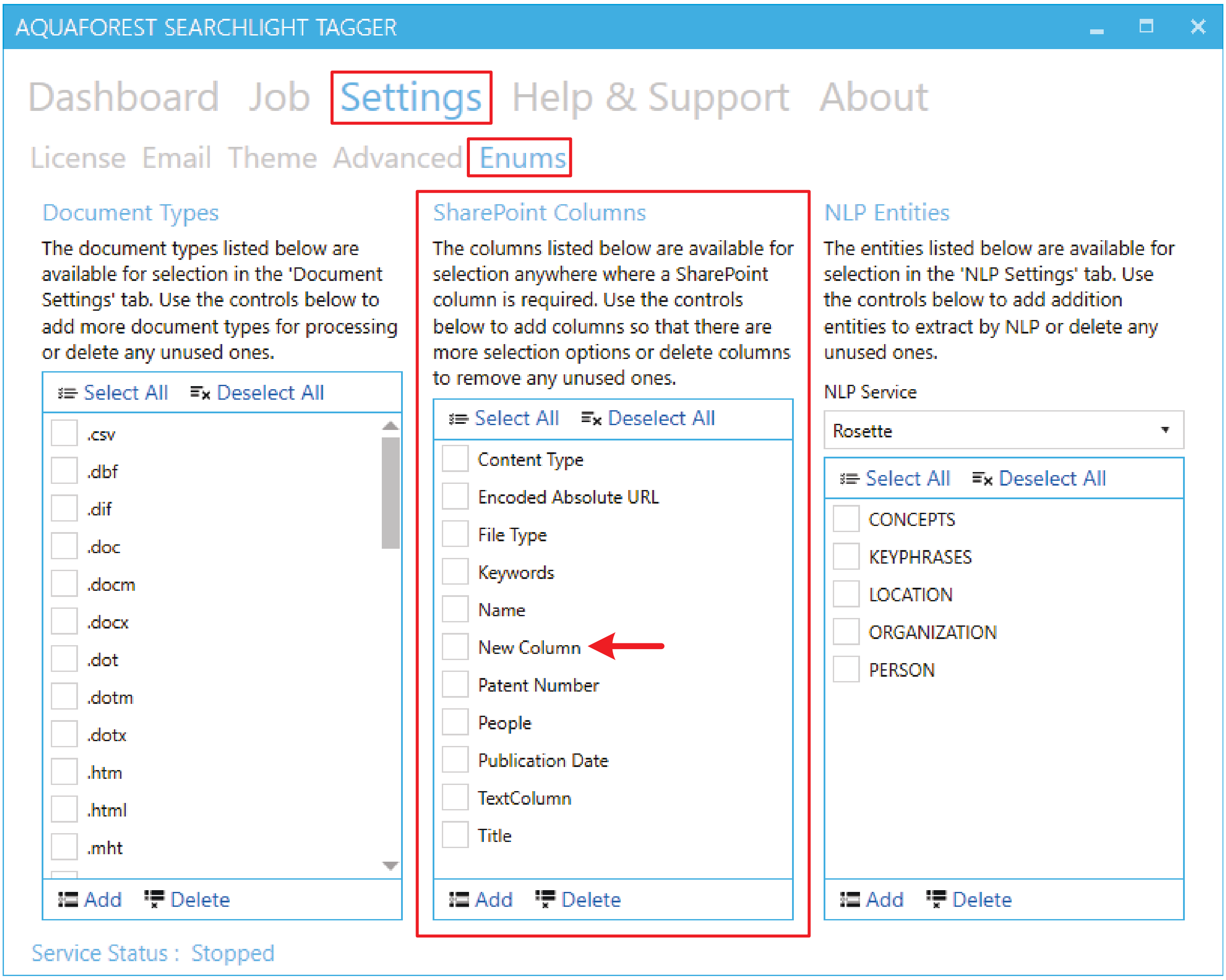


NOTE: SharePoint column names are case-sensitive

Once the drop-down menu loses focus (e.g. you click on another control), the new SharePoint column becomes available for future selection on the current setting and once saved, it is available on all settings where a SharePoint column name is required.



Another way to add SharePoint columns that are not present in the drop-down menu is to go to **Settings** > **Enums** tab.



AQUAFOREST SEARCHLIGHT TAGGER

Dashboard Job Settings Help & Support About

License Email Theme Advanced Enums

Document Types

The document types listed below are available for selection in the 'Document Settings' tab. Use the controls below to add more document types for processing or delete any unused ones.

Select All Deselect All

- □ .csv
- □ .dbf
- □ .dif
- □ .doc
- □ .docm
- □ .docx
- □ .dot
- □ .dotm
- □ .dotx
- □ .htm
- □ .html
- □ .mht

Add Delete

SharePoint Columns

The columns listed below are available for selection anywhere where a SharePoint column is required. Use the controls below to add columns so that there are more selection options or delete columns to remove any unused ones.

Select All Deselect All

- □ Content Type
- □ Encoded Absolute URL
- □ File Type
- □ Keywords
- □ Name
- □ New Column
- □ Patent Number
- □ People
- □ Publication Date
- □ TextColumn
- □ Title

Add Delete

NLP Entities

The entities listed below are available for selection in the 'NLP Settings' tab. Use the controls below to add addition entities to extract by NLP or delete any unused ones.

NLP Service

Rosette

Select All Deselect All

- □ CONCEPTS
- □ KEYPHRASES
- □ LOCATION
- □ ORGANIZATION
- □ PERSON

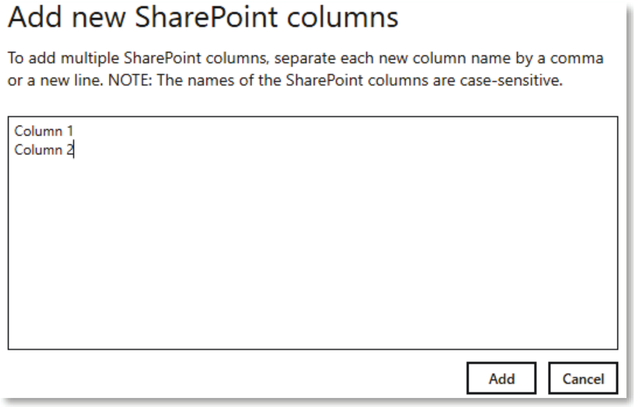
Add Delete

Service Status : Stopped

(The red arrow shows the column that was added by typing it in the drop down menu)

1 Click on the Add button

2 A popup dialog will appear. Enter the name(s) of the SharePoint column(s).



Add new SharePoint columns

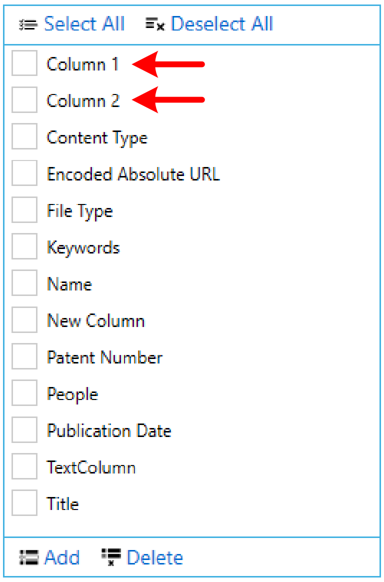
To add multiple SharePoint columns, separate each new column name by a comma or a new line. NOTE: The names of the SharePoint columns are case-sensitive.

Column 1
Column 2

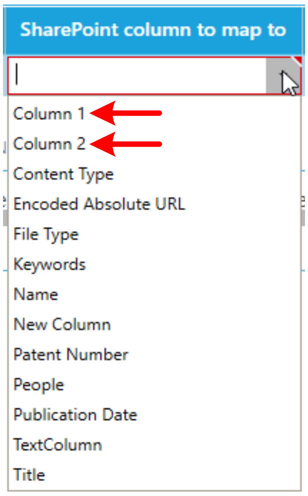
Add Cancel

With this method, you can add multiple SharePoint columns in one go. Separate each new SharePoint column name by a comma or a new line.

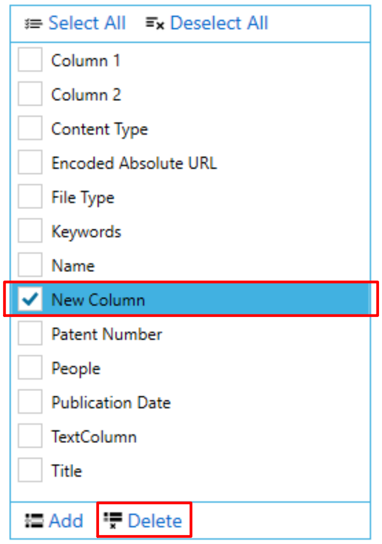
Click the **Add** button after adding all the SharePoint columns.



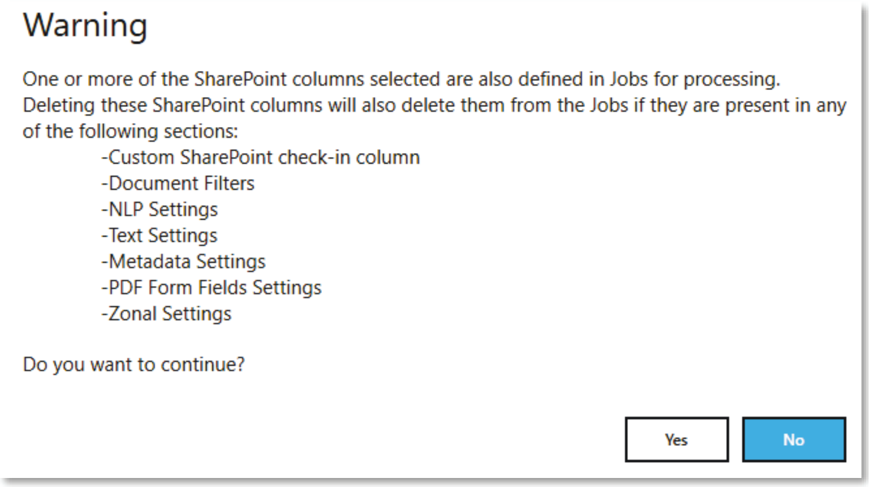
Now these columns will be available on all drop-down menus where a SharePoint column is required.



You can also delete any unused SharePoint columns. Select the column(s) you want to delete and click on the Delete button



However, make sure that the columns you are deleting are not defined in a setting. If it is, you will get a warning message:



Warning

One or more of the SharePoint columns selected are also defined in Jobs for processing. Deleting these SharePoint columns will also delete them from the Jobs if they are present in any of the following sections:

- -Custom SharePoint check-in column
- -Document Filters
- -NLP Settings
- -Text Settings
- -Metadata Settings
- -PDF Form Fields Settings
- -Zonal Settings

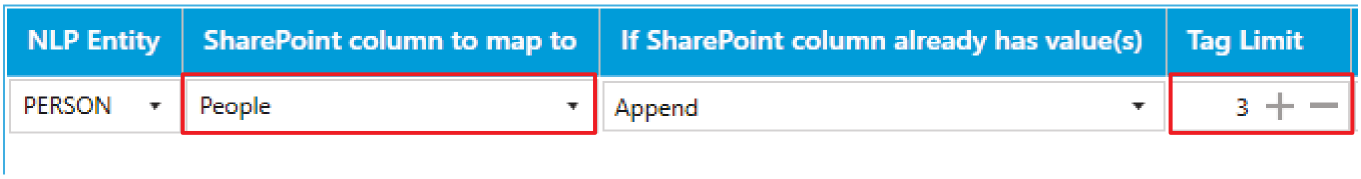
Do you want to continue?

5.5 Tag Limits

Tag limits enable you to restrict the number of metadata that is added (tagged) to a specific SharePoint column.

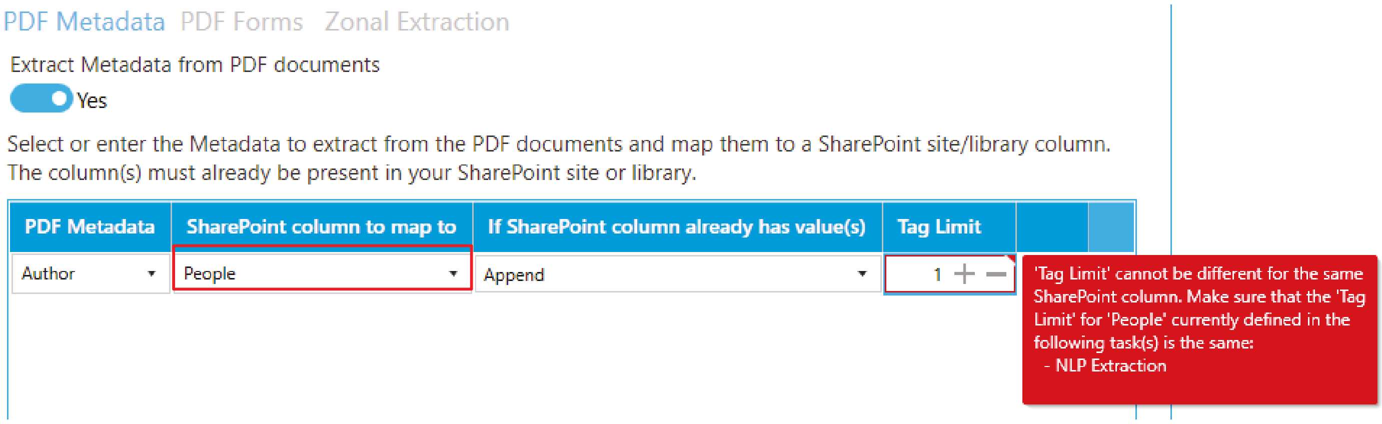
Tag limits are shared among extraction tasks. For instance, say you have a Job where you enabled NLP extraction and PDF metadata extraction.

For the NLP extraction, you have the following settings:



NLP Entity	SharePoint column to map to	If SharePoint column already has value(s)	Tag Limit
PERSON	People	Append	3

For the PDF Metadata settings, you will not be able to set a different **Tag Limit** for the same SharePoint column:



Extract Metadata from PDF documents

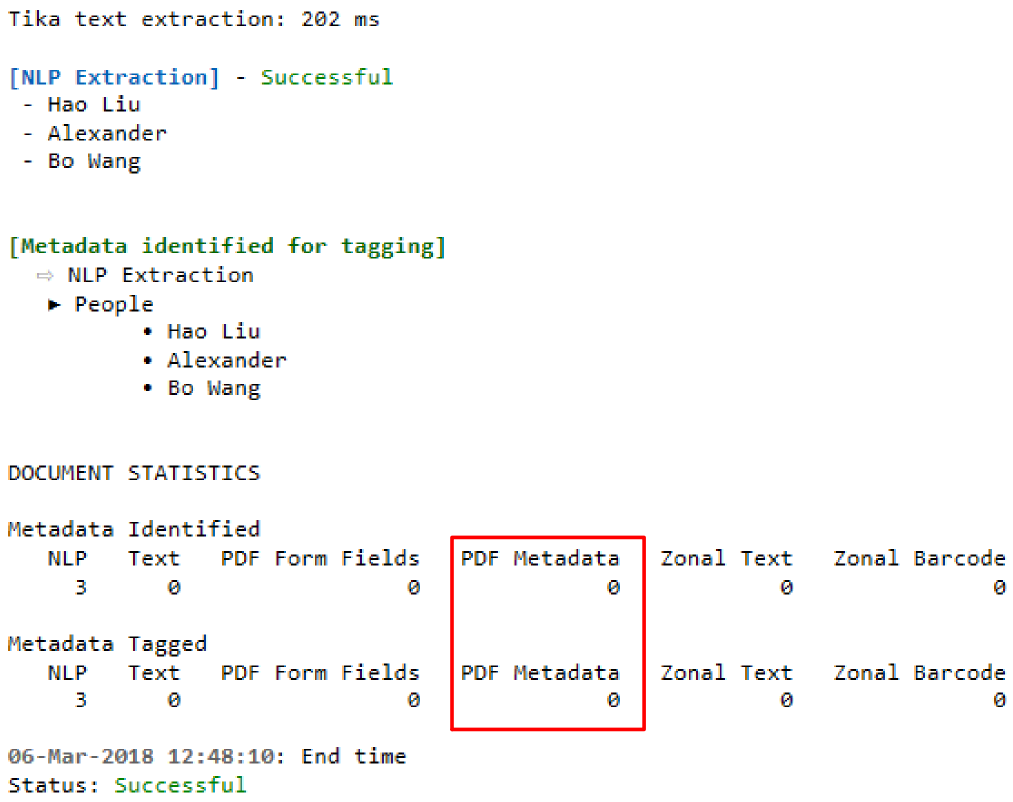
Yes

Select or enter the Metadata to extract from the PDF documents and map them to a SharePoint site/library column. The column(s) must already be present in your SharePoint site or library.

PDF Metadata	SharePoint column to map to	If SharePoint column already has value(s)	Tag Limit
Author	People	Append	1

'Tag Limit' cannot be different for the same SharePoint column. Make sure that the 'Tag Limit' for 'People' currently defined in the following task(s) is the same:
- NLP Extraction

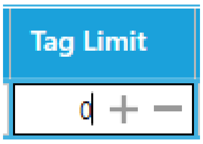
If one of the extraction task hits the **Tag Limit**, the other extraction tasks will be skipped. Using the above example and setting both tag limits to 3, we get the following output:



```
Tika text extraction: 202 ms

[NLP Extraction] - Successful
 - Hao Liu
 - Alexander
 - Bo Wang


[Metadata identified for tagging]
   ⇨ NLP Extraction
   ► People
          • Hao Liu
          • Alexander
          • Bo Wang


DOCUMENT STATISTICS

Metadata Identified
   NLP    Text    PDF Form Fields    PDF Metadata    Zonal Text    Zonal Barcode
     3      0                  0               0             0                0

Metadata Tagged
   NLP    Text    PDF Form Fields    PDF Metadata    Zonal Text    Zonal Barcode
     3      0                  0               0             0                0

06-Mar-2018 12:48:10: End time
Status: Successful
```

As you can see, since the NLP extraction already extracted 3 metadata for the People SharePoint column, the PDF metadata extraction is not performed. See section 4.10.1.1 to see how to read log outputs.

If you do not want to limit any tagging, set the **Tag Limit** to ‘**0**’.

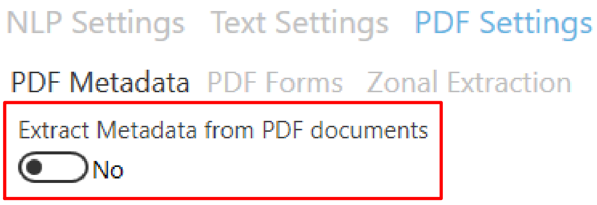


Tag Limit
0

For the above example, since both tag limits have been set to ‘3’, you will need to set one of the extraction task to false before setting both of them to ‘0’.

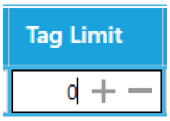
Example:

1. Disable PDF Metadata extraction



Extract Metadata from PDF documents
No

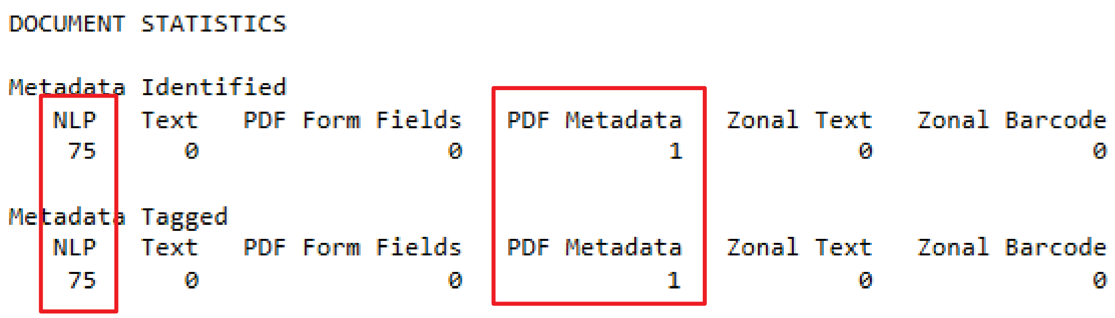
2. Go to NLP Settings and set the **Tag Limit** to ‘0’ and click **Save**.



Tag Limit
0

3. Go back PDF Settings and enable PDF Metadata extraction
4. Set its **Tag Limit** to ‘0’ and click **Save**.

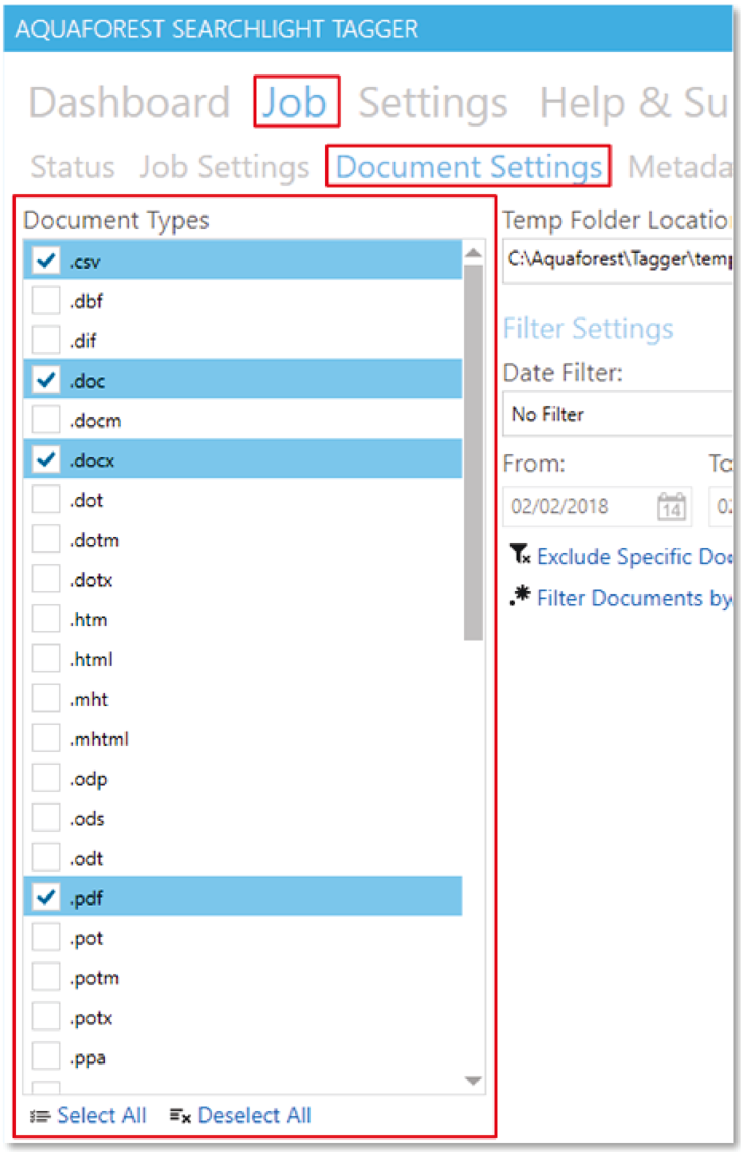
If you run the job, you will see both extraction tasks are performed.



```
DOCUMENT STATISTICS

Metadata Identified
   NLP    Text   PDF Form Fields   PDF Metadata   Zonal Text   Zonal Barcode
    75       0                 0              1            0               0

Metadata Tagged
   NLP    Text   PDF Form Fields   PDF Metadata   Zonal Text   Zonal Barcode
    75       0                 0              1            0               0
```

5.6 Document Types

Tagger can support more document formats that is available by default for selection under **Job** > **Document Settings**.



Document Types

- ☑ .csv
- □ .dbf
- □ .dif
- ☑ .doc
- □ .docm
- ☑ .docx
- □ .dot
- □ .dotm
- □ .dotx
- □ .htm
- □ .html
- □ .mht
- □ .mhtml
- □ .odp
- □ .ods
- □ .odt
- ☑ .pdf
- □ .pot
- □ .potm
- □ .potx
- □ .ppa

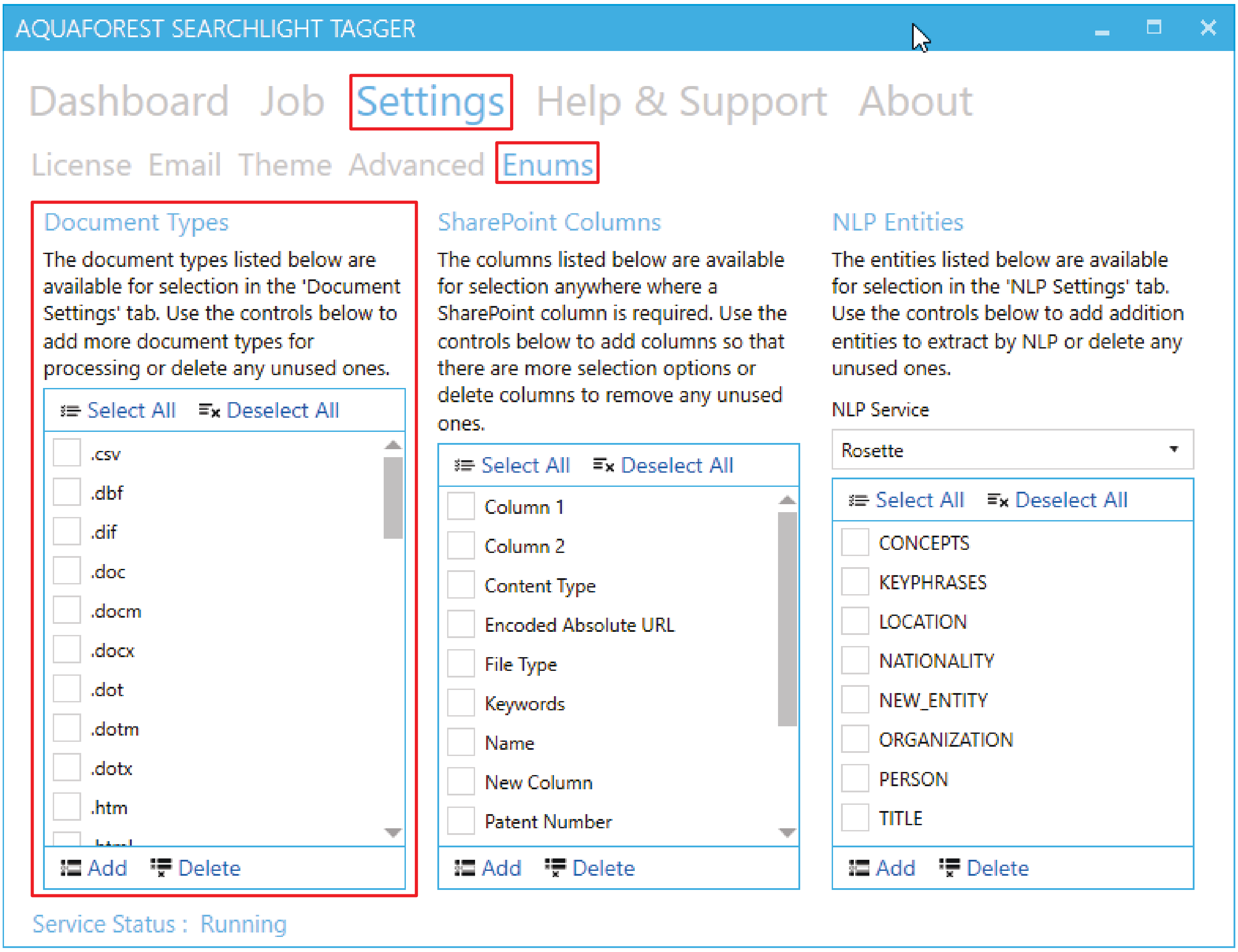
Temp Folder Locatio

C:\Aquaforest\Tagger\tem

Date Filter: No Filter

From: 02/02/2018 To: 0

The document types available for selection are controlled by the **Document Types** section in **Settings** > **Enums** tab.



AQUAFOREST SEARCHLIGHT TAGGER

Dashboard Job Settings Help & Support About

License Email Theme Advanced Enums

Document Types

The document types listed below are available for selection in the 'Document Settings' tab. Use the controls below to add more document types for processing or delete any unused ones.

Select All Deselect All

- □ .csv
- □ .dbf
- □ .dif
- □ .doc
- □ .docm
- □ .docx
- □ .dot
- □ .dotm
- □ .dotx
- □ .htm

Add Delete

SharePoint Columns

The columns listed below are available for selection anywhere where a SharePoint column is required. Use the controls below to add columns so that there are more selection options or delete columns to remove any unused ones.

Select All Deselect All

- □ Column 1
- □ Column 2
- □ Content Type
- □ Encoded Absolute URL
- □ File Type
- □ Keywords
- □ Name
- □ New Column
- □ Patent Number

Add Delete

NLP Entities

The entities listed below are available for selection in the 'NLP Settings' tab. Use the controls below to add addition entities to extract by NLP or delete any unused ones.

NLP Service: Rosette

Select All Deselect All

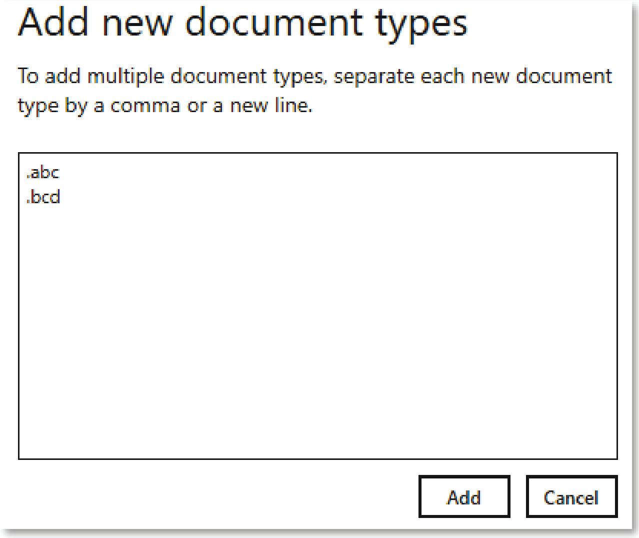
- □ CONCEPTS
- □ KEYPHRASES
- □ LOCATION
- □ NATIONALITY
- □ NEW_ENTITY
- □ ORGANIZATION
- □ PERSON
- □ TITLE

Add Delete

Service Status : Running

If you want to process a particular document format that is not available for selection, you can add it as follows:

1. Click on the Add button
2. A popup dialog will appear. Enter the document type(s) preceded by a dot (.).



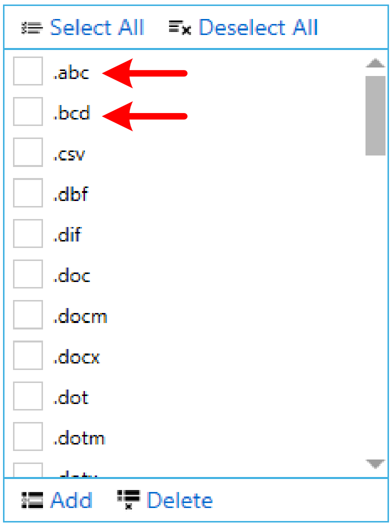
Add new document types

To add multiple document types, separate each new document type by a comma or a new line.

```
.abc
.bcd
```

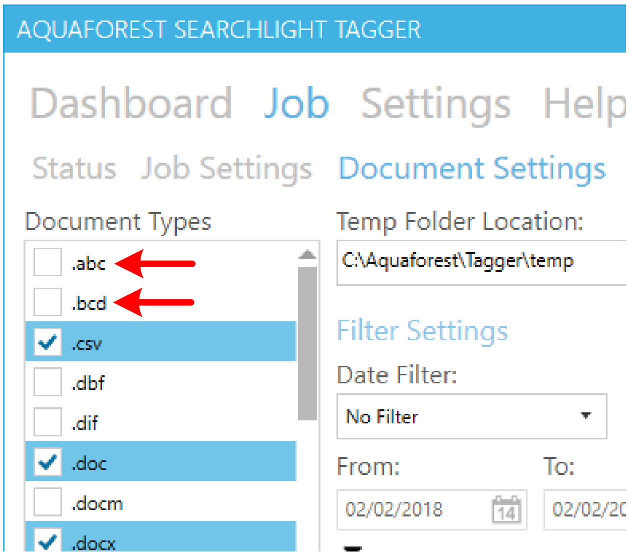
Add Cancel

You can add multiple document types by separating each one by a comma or a new line. Click the **Add** button after specifying all the new document types to make them available for selection.



□ .abc
□ .bcd
□ .csv
□ .dbf
□ .dif
□ .doc
□ .docm
□ .docx
□ .dot
□ .dotm

Now these document types will be available for selection under **Job** > **Document Settings**.



Document Types

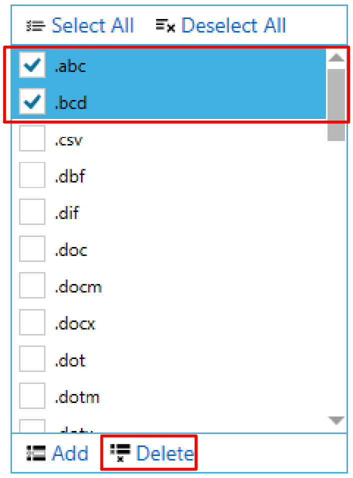
□ .abc
□ .bcd
☑ .csv
□ .dbf
□ .dif
☑ .doc
□ .docm
☑ .docx

Temp Folder Location: C:\Aquaforest\Tagger\temp

Date Filter: No Filter

From: 02/02/2018 To: 02/02/20

You can also delete any unused document types. Select the document type(s) you want to delete and click on the Delete button



☑ .abc
☑ .bcd
□ .csv
□ .dbf
□ .dif
□ .doc
□ .docm
□ .docx
□ .dot
□ .dotm

5.7 Running Searchlight Tagger with Searchlight OCR

In order to extract metadata from PDF documents using Entity Extraction, Taxonomy Matching and Zonal Text Extraction, the PDF documents must be text searchable in the first place. If they are image-only, these extraction tasks will fail because there will be no text to extract and process.

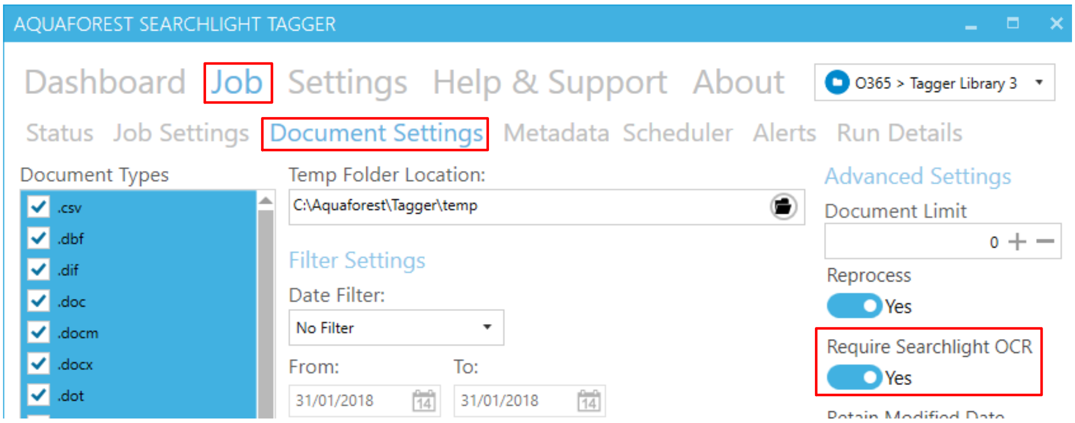
To overcome this issue, you can use Searchlight Tagger in conjunction with Searchlight OCR to ensure that PDF documents are fully text searchable before Tagger attempts to process them.

For this to work:

1. Both Searchlight Tagger and Searchlight OCR must be installed on the same machine.

2. You will need to create a library in Searchlight OCR that points to the site collection, site or library that you are processing in Tagger and schedule it to run before Tagger.

3. In Tagger, go to **Job** > **Document Settings** and set **Require Searchlight OCR** to 'Yes'.



Document Types
☑ .csv
☑ .dbf
☑ .dif
☑ .doc
☑ .docm
☑ .docx
☑ .dot

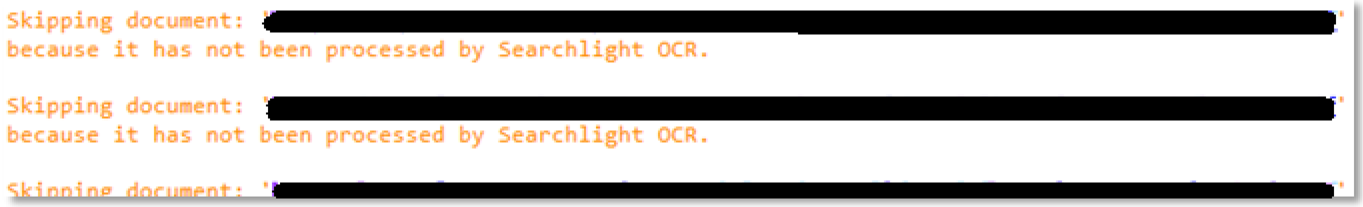
Temp Folder Location: C:\Aquaforest\Tagger\temp

Filter Settings
Date Filter: No Filter
From: 31/01/2018
To: 31/01/2018

Advanced Settings
Document Limit: 0
Reprocess: Yes
Require Searchlight OCR: Yes
Retain Modified Date

4. Schedule it to run after Searchlight OCR.

Tagger will automatically identify where Searchlight OCR is installed and query its database to see the documents that have been processed by Searchlight OCR. If Tagger encounters a document that has not been processed by Searchlight OCR, it will skip the document and display the following warning message in the log file.



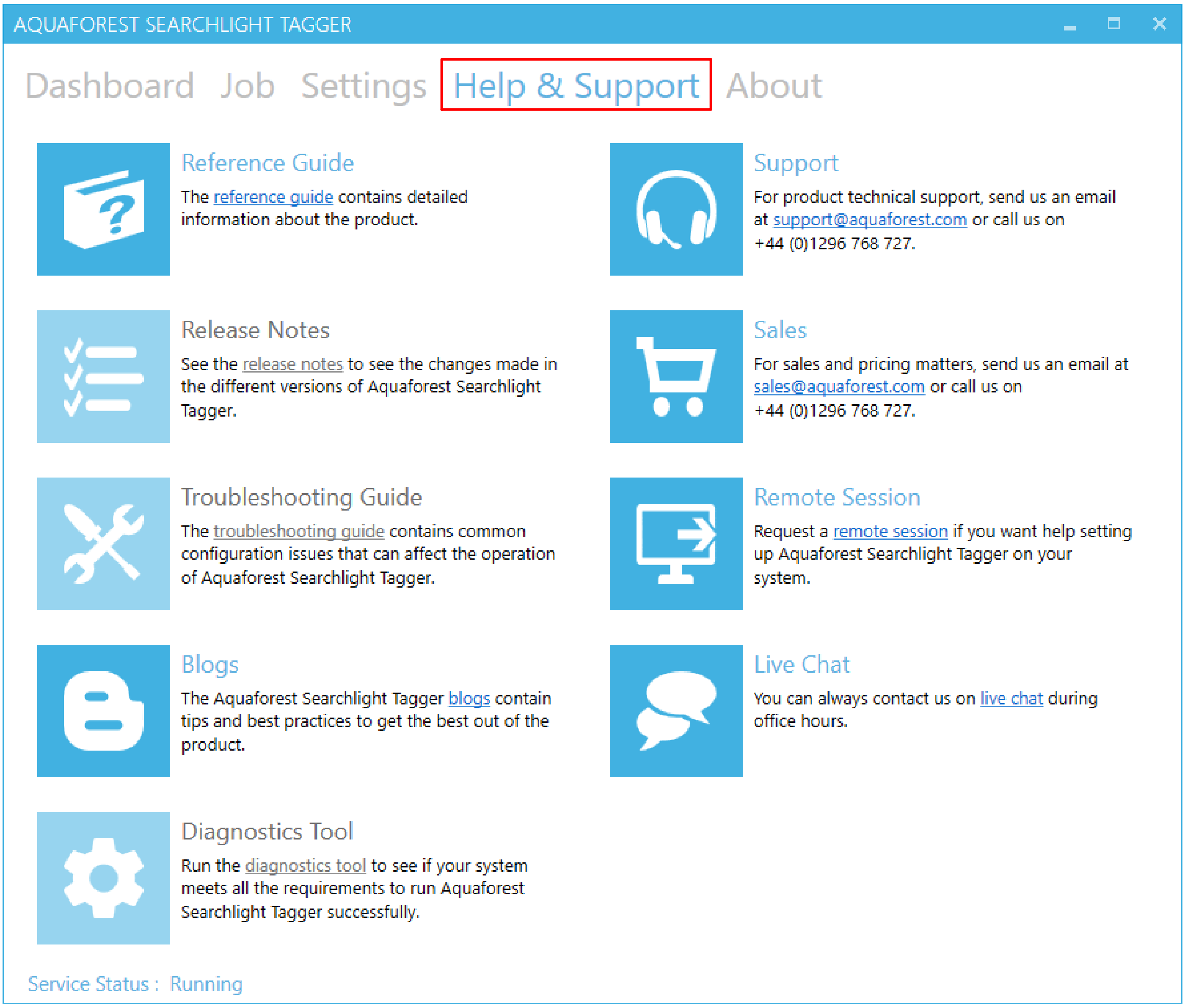
```
Skipping document: '[redacted]'
because it has not been processed by Searchlight OCR.

Skipping document: '[redacted]'
because it has not been processed by Searchlight OCR.

Skipping document: '[redacted]'
```

Tagger will keep skipping the document until it is processed by Searchlight OCR.

If Searchlight OCR is not installed, Tagger will not work unless you set **Require Searchlight OCR** (see step 3 above) to 'No'.

5.8 Help & Support

The Help & Support page is the starting point for help with Aquaforest Searchlight. It provides resources such as the reference guide, release notes and online blogs. It also provides the generic support email address, which should be used in the first instance when reporting an issue or any queries.



Reference Guide

The reference guide contains detailed information about the product.

Support

For product technical support, send us an email at support@aquaforest.com or call us on +44 (0)1296 768 727.

Release Notes

See the release notes to see the changes made in the different versions of Aquaforest Searchlight Tagger.

Sales

For sales and pricing matters, send us an email at sales@aquaforest.com or call us on +44 (0)1296 768 727.

Troubleshooting Guide

The troubleshooting guide contains common configuration issues that can affect the operation of Aquaforest Searchlight Tagger.

Remote Session

Request a remote session if you want help setting up Aquaforest Searchlight Tagger on your system.

Blogs

The Aquaforest Searchlight Tagger blogs contain tips and best practices to get the best out of the product.

Live Chat

You can always contact us on live chat during office hours.

Diagnostics Tool

Run the diagnostics tool to see if your system meets all the requirements to run Aquaforest Searchlight Tagger successfully.

Service Status : Running

6 Acknowledgements

This product makes use of a number of Open Source components, which are included in binary form. The appropriate acknowledgements and copyright notices are given below.

Name	Homepage
AutoMapper	Homepage | GitHub
AvalonEdit	Homepage | GitHub
BitMiracle.LibTiff.NET	Homepage | GitHub
BouncyCastle.Crypto	Homepage
Common.Logging	Homepage
CompareNETObjects	GitHub
CronExpressionDescriptor	Homepage | GitHub
Extended.Wpf.Toolkit	Homepage | GitHub
IKVM.NET	Homepage | Sourceforge
Log4Net	Homepage
Lucene.Net	Homepage
MahApps MahApps.Metro MahApps.Metro.IconPacks	Homepage GitHub GitHub
Microsoft.WindowsAPICodePack.Core	Homepage
Microsoft.WindowsAPICodePack.Shell	Homepage
Newtonsoft.Json	Homepage | GitHub
PDFBox	Homepage
Quartz	Homepage | GitHub
System.Data.SQLite	Homepage
Tika	Homepage
ZXing.Net	Homepage | GitHub

- button on the Dashboard instead.