Office 365 SharePoint

Automated Metadata Tagging for SharePoint and Office 365

By Agnes Molnar, Search Explained

www.aquaforest.com

This white paper provides an overview of metadata and tagging capabilities of SharePoint and Office 365 as well as useful instructions, and practical advice regarding how to extend and improve the out-of-the-box features. The paper has been prepared by Agnes Molnar, Office Servers and Services MVP, Founder and Principal Consultant of Search Explained.

Contents

Introduction

Every day, everywhere, information workers work with a huge volume of content: they create as well as consume. They have to be able to **find** the document they need. They have to **decide** if a newly discovered item is the one they need or it's better to search further. They have to be able to **use** the content the way they want. And they have to **make business-critical decisions** fast, based on the content they find.

And this is not as simple as it sounds to be.

There are many different things that have to be in place and fit together in order to make content usability and findability good and successful.

If you can get your information architecture and metadata right, your users will be able to find and use content they actually need fast and efficiently.

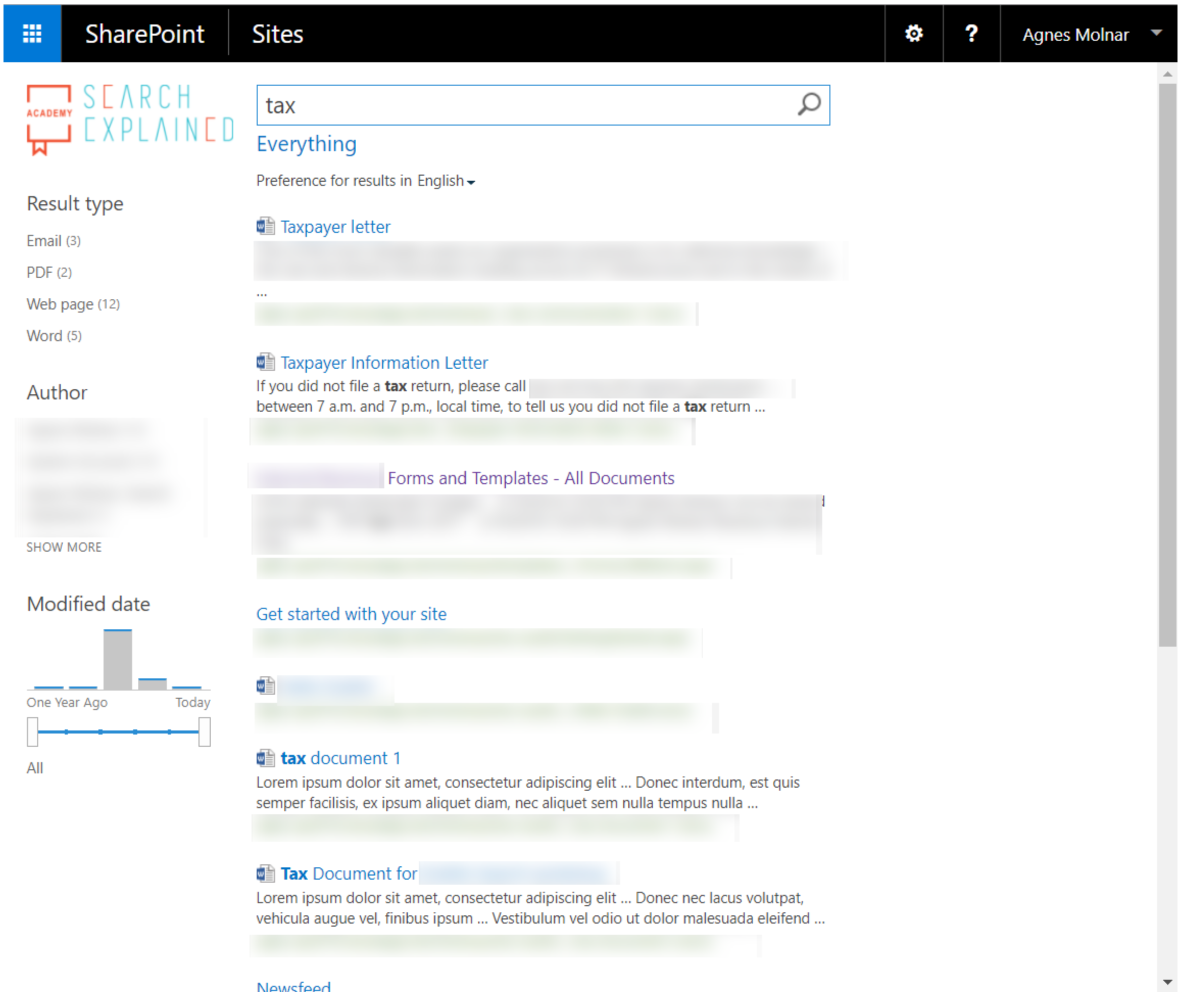
However, if you don't have good quality metadata on your content, the chance of your Search being good is tiny. Your content without metadata is no more than files stored in a network drive. Findability is very poor. Usually, the only way to get to a document is to navigate there. Over time, due to the lack of findability, your users will end up adding the same contents over and over, resulting in exponential growth of duplicate contents. Eventually, you will end up having a content silo – with close-to-zero findability of your documents.

Another common issue is when you have metadata on the content, but it's bad and inconsistent. This might happen for several reasons:

- your users do not have the required knowledge;
- your users are not sure how to create good metadata, or how to use the forms the right way;
- or they are not motivated to spend even a couple of minutes to fill in the properties;
- or a combination of the reasons above.

The result is incorrect, inconsistent and messy metadata that makes content usability and findability even worse. Bad metadata is misleading. Inconsistent metadata is hard to track and correct. There is no way to overcome these issues, other than fixing your metadata itself.

And regardless of how much time, money, knowledge, and expertise you invest into your Search application, it will give you more headache than help. Even Search cannot help, because it cannot rely on anything else but the out-of-the-box configuration:



Taxpayer letter

...

Taxpayer Information Letter

If you did not file a **tax** return, please call
between 7 a.m. and 7 p.m., local time, to tell us you did not file a **tax** return ...

Forms and Templates - All Documents

Get started with your site

tax document 1

Lorem ipsum dolor sit amet, consectetur adipiscing elit ... Donec interdum, est quis semper facilisis, ex ipsum aliquet diam, nec aliquet sem nulla tempus nulla ...

Tax Document for

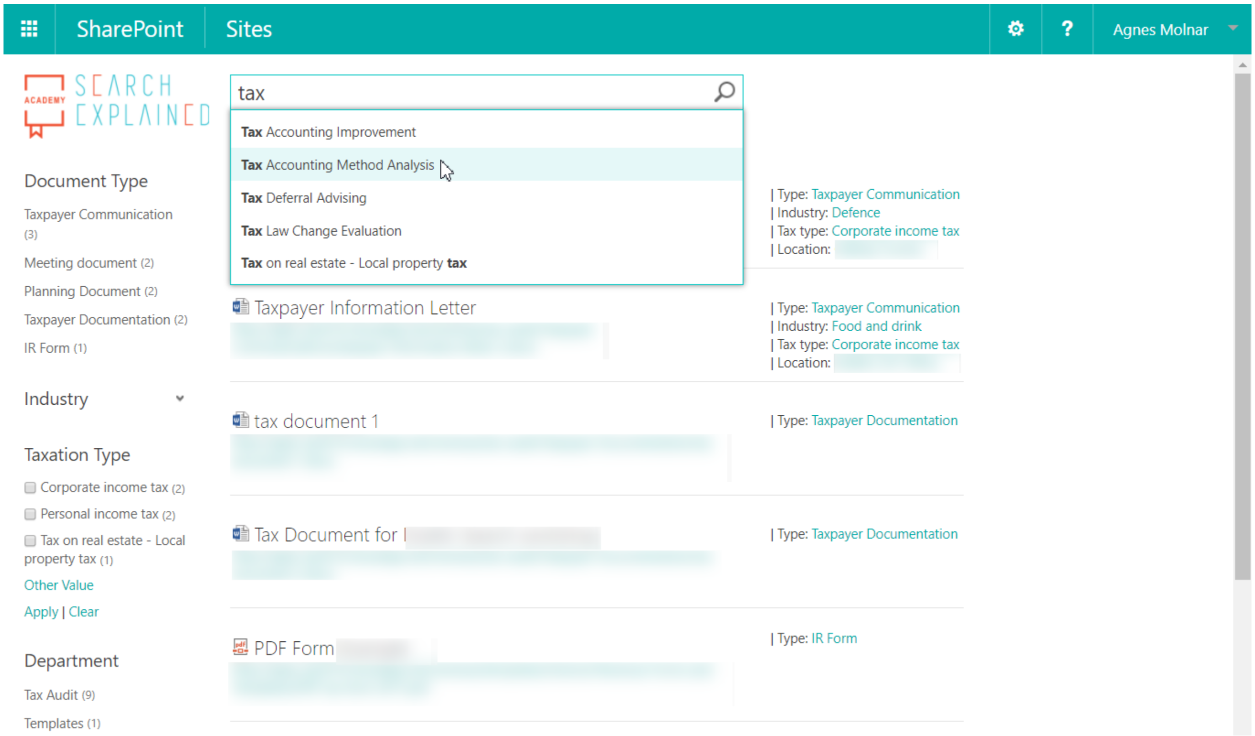
Lorem ipsum dolor sit amet, consectetur adipiscing elit ... Donec nec lacus volutpat, vehicula augue vel, finibus ipsum ... Vestibulum vel odio ut dolor malesuada eleifend ...

Newsfeed

Therefore, having good metadata on your content is essential. The benefits are obvious:

- Improved usability and findability of content;
- Improved search applications;
- Less time spent with not finding the content;
- Better overall user satisfaction;
- Etc.

With good quality metadata not only the usability, but also *findability* of your content skyrockets. The result is: happy employees who can get their jobs done much faster and easier:



tax

Tax Accounting Improvement
Tax Accounting Method Analysis
Tax Deferral Advising
Tax Law Change Evaluation
Tax on real estate - Local property **tax**

| Type: Taxpayer Communication
| Industry: Defence
| Tax type: Corporate income tax
| Location: [redacted]

Taxpayer Information Letter [redacted]

| Type: Taxpayer Communication
| Industry: Food and drink
| Tax type: Corporate income tax
| Location: [redacted]

tax document 1 [redacted]

| Type: Taxpayer Documentation

Tax Document for [redacted]

| Type: Taxpayer Documentation

PDF Form [redacted]

| Type: IR Form

Types of Metadata

First, let me explain what types of metadata we can define in SharePoint, to provide you a lay of the land.

In most cases, we use unmanaged metadata in SharePoint:

- Single line text
- Multi-line text
- Number
- Date / time
- Etc.

In case of these types of metadata, users can enter their own values; therefore, the set of values might be very broad, uncontrolled and inconsistent. The users are free to use different forms and synonyms without any (external) control. We can set up rules and governance practices about what values should be used, but it's everyone's own responsibility to follow these guidelines.

In other cases, metadata is managed by "metadata owners" or taxonomists: a group of users who are responsible for creating, maintain and curate the metadata as part of the organization's knowledge management system:

> *"The managed metadata environment represents the architectural components, people and processes that are required to properly and systematically gather, retain and disseminate metadata throughout the enterprise."*
>
> (from "Building and Managing the Metadata Repository"
> by David Marco, J. Wiley, 2000)

Using managed metadata has several benefits, of course:

- controlled, consistent set of metadata values;
- rules and governance practices provide the quality of managed metadata;
- simple data discovery;
- increased confidence;
- rely and usage of staff knowledge regarding to business rules and definitions;
- improved cooperation between business and IT.

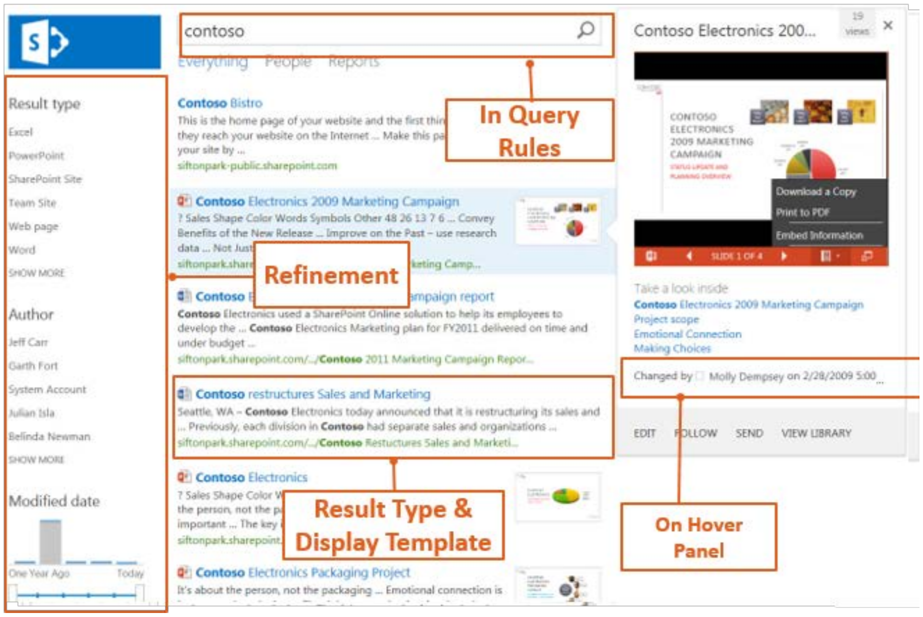
In SharePoint, there are several ways to store managed data types:

- Choice
- Lookup
- Managed Metadata

Of course, the complexity and typical usage of these vary. The one that provides the biggest complexity, the richest features, and highest flexibility is Managed Metadata, provided by the Term Store.

Quality of metadata

Storing metadata is not enough though. There are many different factors that affect the quality of metadata, and as a result, the quality of usability and findability of the content too. Users find content that has good quality metadata much easier. Navigation, classic search, filtering, etc. – everything is much easier. The diagram below demonstrates how **everything** is driven by metadata on a classic search page.



Of course, besides this, search can drive navigation, displaying aggregated and/or filtered content anywhere

If the metadata is managed, it's easier to control what values the users enter – but harder to maintain the environment.

If the metadata is unmanaged, there's no maintenance effort needed (or it's minimal) – but the users might enter unpredictable and inconsistent values.

The decision that you have to make is always a trade-off. The primary consideration is the quality of content metadata. – Why? Because the quality of content usability, as well as findability, depends on it. Why would you create a document if you don't want it to be found and used?

Automated Generation of Metadata

We, human beings, tend to make mistakes, there is no exception. Some users are really overwhelmed with their tasks and don't have enough time and/or energy to add the proper metadata to every document they create. Others lack the education and don't know how to do it the right way. There are also users who are not motivated enough.

Auto-tagging solutions can provide an enormous help here. They generate metadata automatically, with zero or at least minimal end-user interaction. With these tools, you can avoid all the human mistakes described above.

The big question is always how to make these tools work? – The bad news is they are not "smart" enough to figure out the users' intent in all the cases, but we can configure "teach" them to do a decent job.

Let me show you the options and supporting features.

Rule-based tagging

The first approach is to define **rules** which describe how and what metadata must be created on the content.

- **Text Rules**: The most common rule tags the documents by comparing their texts with terms in the SharePoint Term Store. If the text of a document matches a term in the Term Store, the relevant term will be automatically added as a tag value to the document. Advanced matching rules may contain regular expressions too.
- **Zonal Rules**: Similar or same-type documents often have the very same layout, like invoices, registers, etc. We can tag documents by extracting text from specific zones in PDF pages.

- **Bar Codes**: Another type of extracting information from a PDF can be to identify and extract barcodes from specific areas of the document, and tag the documents based on the barcode values.

- **PDF Forms and Metadata**: Documents often contain form fields, like customer name, invoice number, product id, etc. These form values can also be extracted and map to a SharePoint column.

As you can see, defining these rules needs planning, preparation, and configuration. We have to define where to extract the information from (zone, barcode, form field), what to extract (values) and how to map these to our managed fields.

To make users' lives easier, we have to invest time and resources in advance.

NLP (Natural Language Processing)-based tagging

According to Wikipedia, Natural Language Processing (NLP) "*is a field of computer science, artificial intelligence concerned with the interactions between computers and human natural) languages, and, in particular, concerned with programming computers* ***to fruitfully process large natural language data.***"

The key is in the last part of the sentence: "to fruitfully process large natural language data."

For example, if you migrate your contents to SharePoint or Office 365 from a network drive (file share), chances are your users won't always fill in the proper, managed or unmanaged fields for any of the reasons described above. In these cases, NLP-based tagging can help you by analyzing the document's text, identifying the possible tags, and adding them to the document automatically.

NLP itself is a very complex topic, I don't want to get into its deep details in this paper. The good news is, however, that you can rely on industry standard NLP services in your applications. Some examples:

- **Rosette**: This service uses "*linguistic analysis, statistical modeling, and machine learning to accurately process unstructured text and names, revealing valuable information and actionable data.*"
  More information: https://www.rosette.com/
- **Open Calais**: Thomson Reuters Open Calais offers an easy and accurate way "*to tag the people, places, companies, facts, and events in your content to increase its value, accessibility, and interoperability.*"
  More information: http://www.opencalais.com/about-open-calais/
- **Microsoft Cognitive Services**: Microsoft's API is also a cloud-based service that includes three main functions: sentiment analysis, key phrase extraction, and language detection.
  More information: https://azure.microsoft.com/en-us/services/cognitive-services/
- **Google Natural Language**: Of course, Google has its own NLP service, too. Google Cloud Natural Language API can be used "*to extract information about people, places, events and much more, mentioned in text documents, news articles or blog posts*", as well as for sentiment analysis.
  More information: https://cloud.google.com/natural-language

As you can see, the list of available NLP services is very impressive. Auto-tagging solutions don't have to write their own NLP / text analytics service anymore; they can rely on these industry standard services.

The benefit of using NLP services, instead of the rule-based tagging features, or beside them, is that we don't have to spend a long time and heavy resources on planning and configuring our rules. However, in this case, please consider that the quality of tagging can be only as good as these services fit to your content corpus.

Warning! *Always do in-depth tests and analysis before you apply any of these to your production data.*

Getting Started and Best Practices

After this introduction of the related technologies, let me show you some practical steps of how to get started with tagging and auto-tagging.

How to decide which metadata columns to define

The first step is always to plan your information architecture. This includes the site structure, lists, and libraries, site columns, Content Types, Managed Metadata term sets, search set-up, etc. – many different dimensions which have to provide an efficient way to store content.

Setting up metadata is only a small piece of the overall Information Architecture, but it's an important one (like everything else, too). One decision has to be where you want to use managed columns and where unmanaged fields.

Some considerations to make:

- If the value set of a field cannot be defined in advance, it has to be an unmanaged field. "Free text" values can be found everywhere: title, description, content outline, subject, etc.
- If the field must be a number or a date, but there's no way to pre-set the values or intervals, it's definitely going to be a numerical or date/time metadata field: date and time of a meeting, number of attendees, price of a flight, etc.
- If the field can have only a few values, and these values don't change over time, choice field is the best way to go with: days of the week (Monday – Sunday), months of the year (January – December), etc.
- If the field can have various, non-hierarchical values, and the goal is easy management of these values by a well-defined group of users, probably you

should consider using lookup fields: name of business units (in case of small businesses where the organizational hierarchy is flat), members of a team, company cars, etc.

- In complex cases, when the value set is hierarchical, you need more control and governance. In these cases, you need to define Managed Metadata Term Sets: locations, organizational hierarchy, document types, tax types, etc.

You can see the decision to be made is not easy. But it has to be made in advance so that we can move forward towards the more advanced features, like auto-tagging.

Pre-requirements

Besides a well-defined metadata structure where the tags can be stored, auto-tagging solutions have another common pre-requirement too: the content's text must be available. In most scenarios, meaningful metadata cannot be extracted from an image (or scanned PDFs, faxes, etc.), so we need to convert them to fully text searchable PDF format using OCR technology, eg. by Aquaforest Searchlight OCR.

Taxonomies requirements

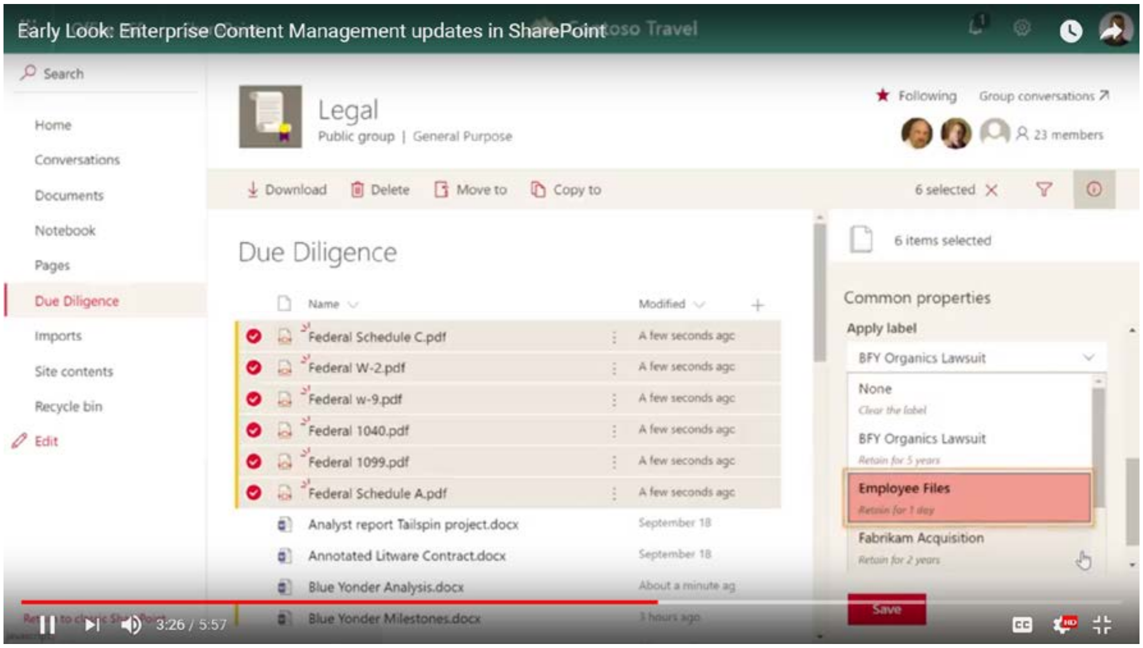
The better taxonomies we want or need, the better we must take care of them. This includes the design, delivery as well as maintenance of these knowledge structures. Combining taxonomies with rule or NLP-based auto-tagging solutions always result in better search and discovery experience.

Taxonomies also help us to have better content governance.

Metadata Features in Office 365

Microsoft has just released an "Early Look" video about "Enterprise Content Management updates in SharePoint". Even though it's about SharePoint Online in

Office 365, it's worth checking the bulk metadata editing and Compliance Label features mentioned in the video. While not directly related to auto-tagging, these features will make users' lives much easier – with the proper planning, governance and compliance settings.



Early Look: Enterprise Content Management updates in SharePoint

Legal
Public group | General Purpose

Due Diligence

Name	Modified
Federal Schedule C.pdf	A few seconds agc
Federal W-2.pdf	A few seconds agc
Federal w-9.pdf	A few seconds agc
Federal 1040.pdf	A few seconds agc
Federal 1099.pdf	A few seconds agc
Federal Schedule A.pdf	A few seconds agc
Analyst report Tailspin project.docx	September 18
Annotated Litware Contract.docx	September 18
Blue Yonder Analysis.docx	About a minute ag
Blue Yonder Milestones.docx	3 hours ago

6 items selected

Common properties

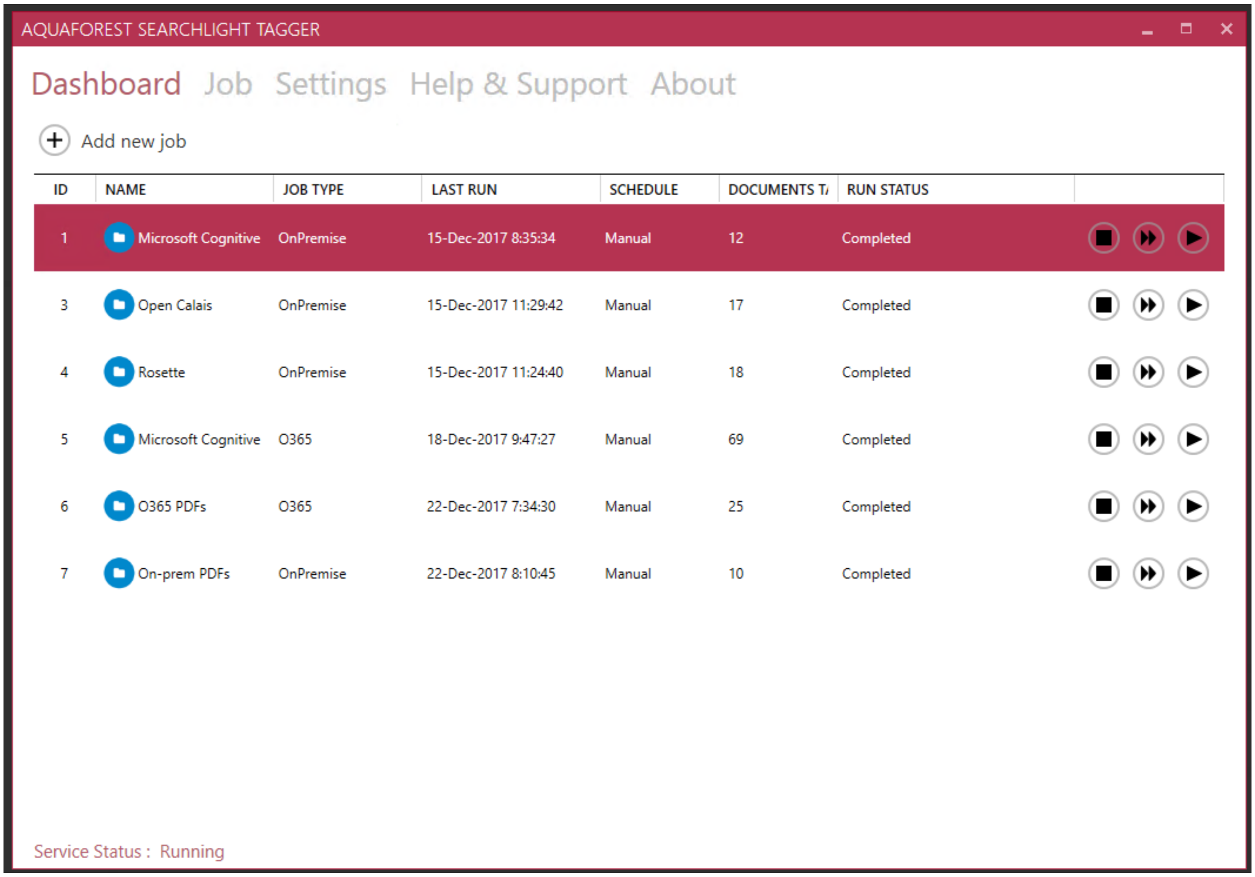
Apply label: BFY Organics Lawsuit

- None – Clear the label
- BFY Organics Lawsuit – Retain for 5 years
- Employee Files – Retain for 1 day
- Fabrikam Acquisition – Retain for 2 years

Aquaforest Searchlight Tagger

Aquaforest Searchlight Tagger is a lightweight but rich tool for auto-tagging. It can be configured to use rules, NLP as well as PDF zones to tag documents – or a combination of these options.

It can work on SharePoint On-Premises (2010, 2013 or 2016) as well as in Office 365.

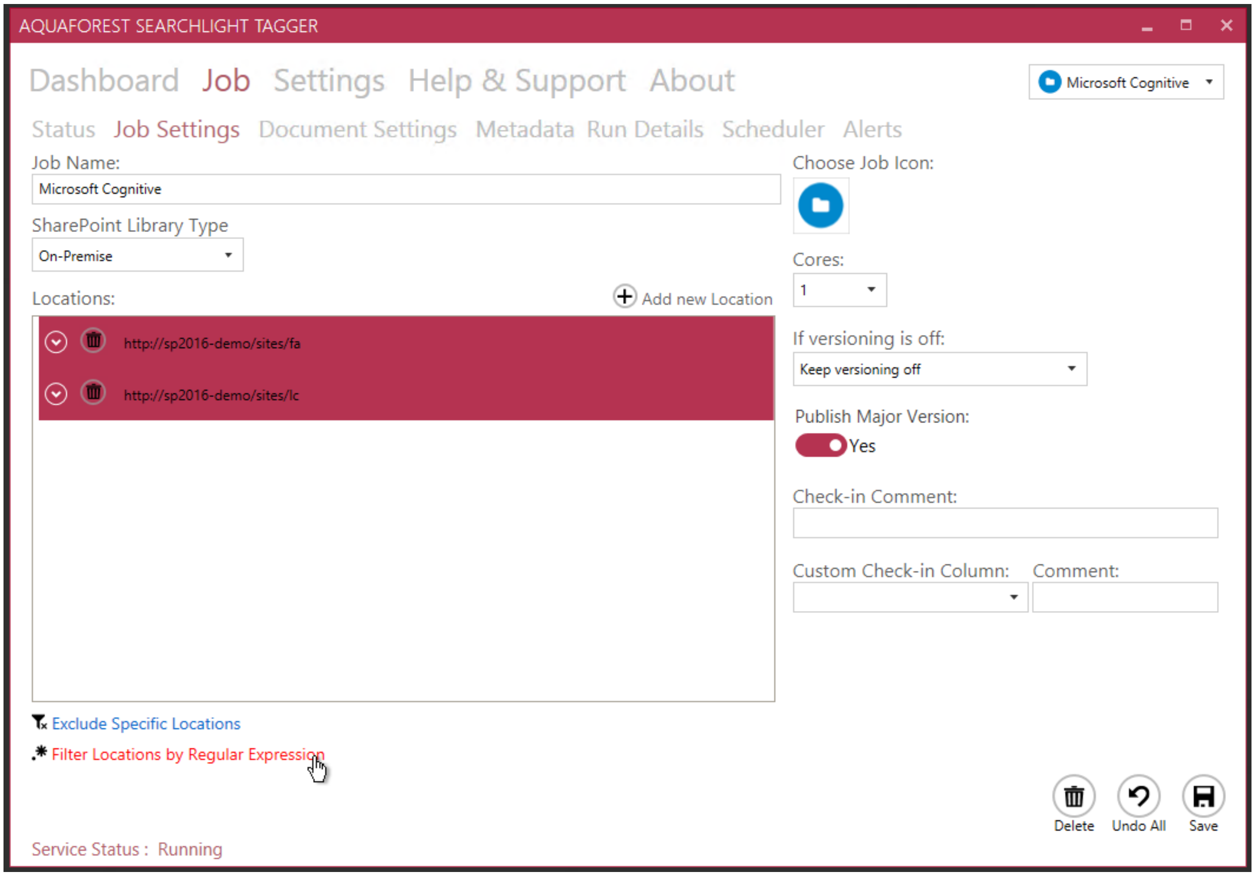


ID	NAME	JOB TYPE	LAST RUN	SCHEDULE	DOCUMENTS T/	RUN STATUS
1	Microsoft Cognitive	OnPremise	15-Dec-2017 8:35:34	Manual	12	Completed
3	Open Calais	OnPremise	15-Dec-2017 11:29:42	Manual	17	Completed
4	Rosette	OnPremise	15-Dec-2017 11:24:40	Manual	18	Completed
5	Microsoft Cognitive	O365	18-Dec-2017 9:47:27	Manual	69	Completed
6	O365 PDFs	O365	22-Dec-2017 7:34:30	Manual	25	Completed
7	On-prem PDFs	OnPremise	22-Dec-2017 8:10:45	Manual	10	Completed

Service Status : Running

When you create a new job, the first thing to define is the type of the deployment: On-Premise or O365. You can add several locations to a single job: Site collections, sites and/or lists and libraries.

You can also filter out specific locations by their URLs as well as by using Regular Expressions.



Job Name: Microsoft Cognitive

SharePoint Library Type: On-Premise

Locations:

- http://sp2016-demo/sites/fa
- http://sp2016-demo/sites/lc

Choose Job Icon:

Cores: 1

If versioning is off: Keep versioning off

Publish Major Version: Yes

Check-in Comment:

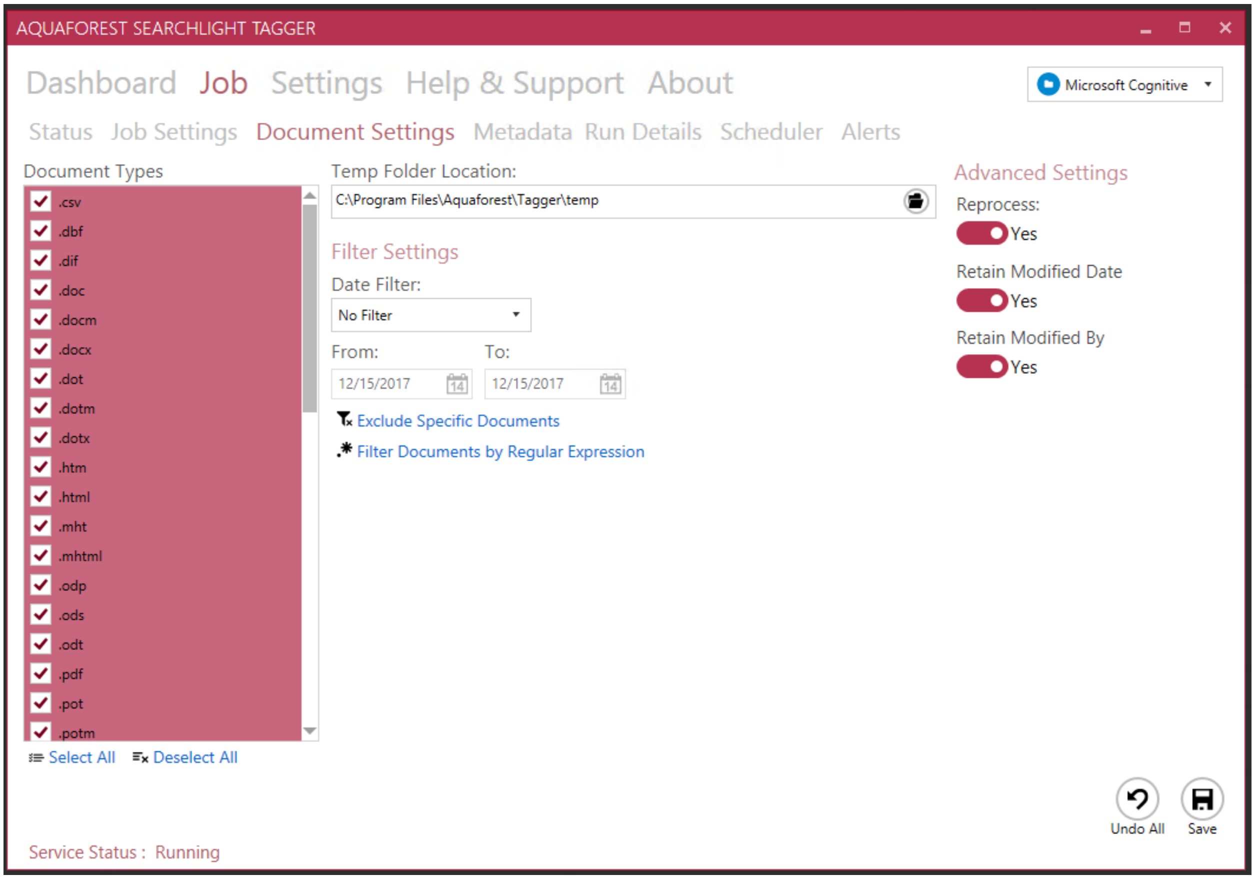
Custom Check-in Column:

Comment:

Service Status : Running

Once the location is defined, next step is to define the Document Settings:

- Document types
- Date filters
- Other filters (by URL or regular expression)
- Retain the “modified date” and “modified by” properties after tagging (always recommended to turn on!)
- Reprocess the documents that already have been tagged – should be set to “Yes” only if you’re doing testing or have changed any configuration settings that could affect the documents that have been tagged earlier.



Document Types

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

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

Filter Settings

Date Filter: No Filter

From: 12/15/2017 To: 12/15/2017

Advanced Settings

Reprocess: Yes

Retain Modified Date: Yes

Retain Modified By: Yes

Service Status : Running

Next step is the "heart" of the whole configuration: metadata settings. You have to be very well prepared, and also, it is always recommended to do detailed tests on a non-production copy of documents.

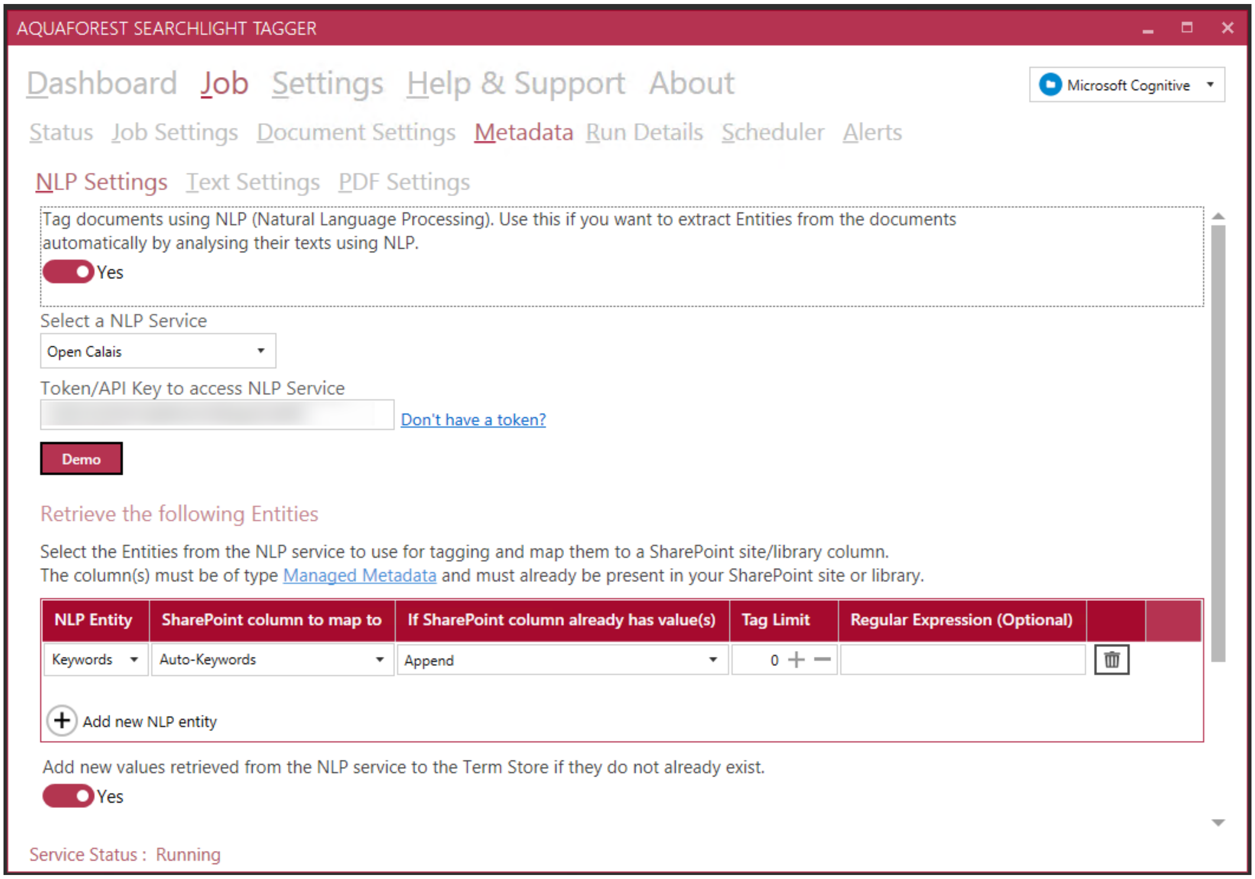
You can define which extraction methods you want to use:

- NLP
- Rule-based (text)
- PDF zones or forms

Each of those pages contains details about the settings, as well as the SharePoint columns (both managed metadata and non-managed metadata) where the identified tags should be stored.

<u>Note</u>: the metadata field must be a Managed Metadata type field. ← only for Text Settings (Taxonomy matching)

<u>Note</u>: To be able to use any of the NLP services, you must have a license with them. For testing and development purposes most of them are free. For production, please consult them for pricing options.



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: Open Calais

Token/API Key to access NLP Service: [redacted] Don't have a token?

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 be of type Managed Metadata and must already be present in your SharePoint site or library.

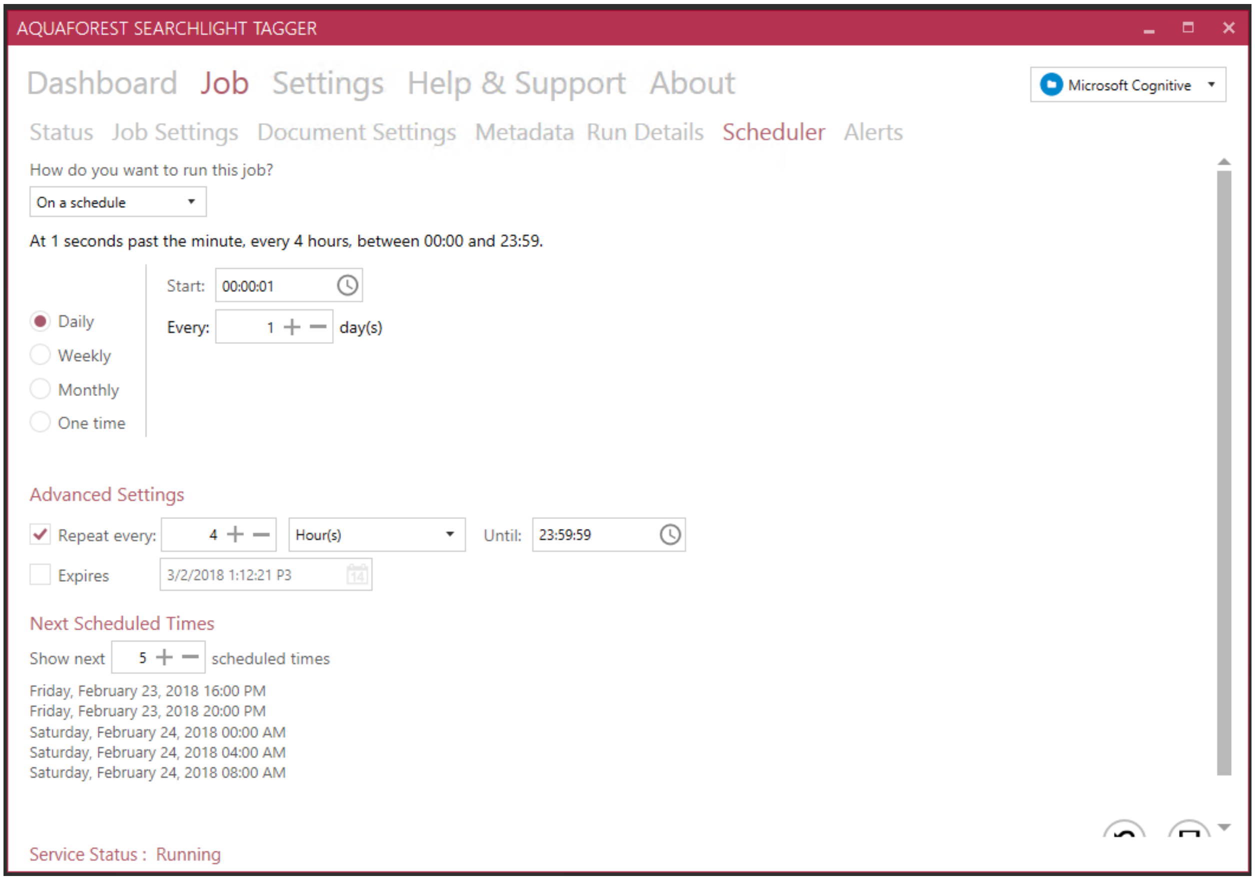
NLP Entity	SharePoint column to map to	If SharePoint column already has value(s)	Tag Limit	Regular Expression (Optional)
Keywords	Auto-Keywords	Append	0	

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

● Yes

Service Status : Running

Last but not least, you have to define the schedule of the job: you can run it manually (always a good option when doing testing or after migrating documents to SharePoint), or you can set it to run regularly every day/week/month.



How do you want to run this job?

On a schedule

At 1 seconds past the minute, every 4 hours, between 00:00 and 23:59.

● Daily
○ Weekly
○ Monthly
○ One time

Start: 00:00:01

Every: 1 day(s)

Advanced Settings

☑ Repeat every: 4 Hour(s) Until: 23:59:59

□ Expires 3/2/2018 1:12:21 P3

Next Scheduled Times

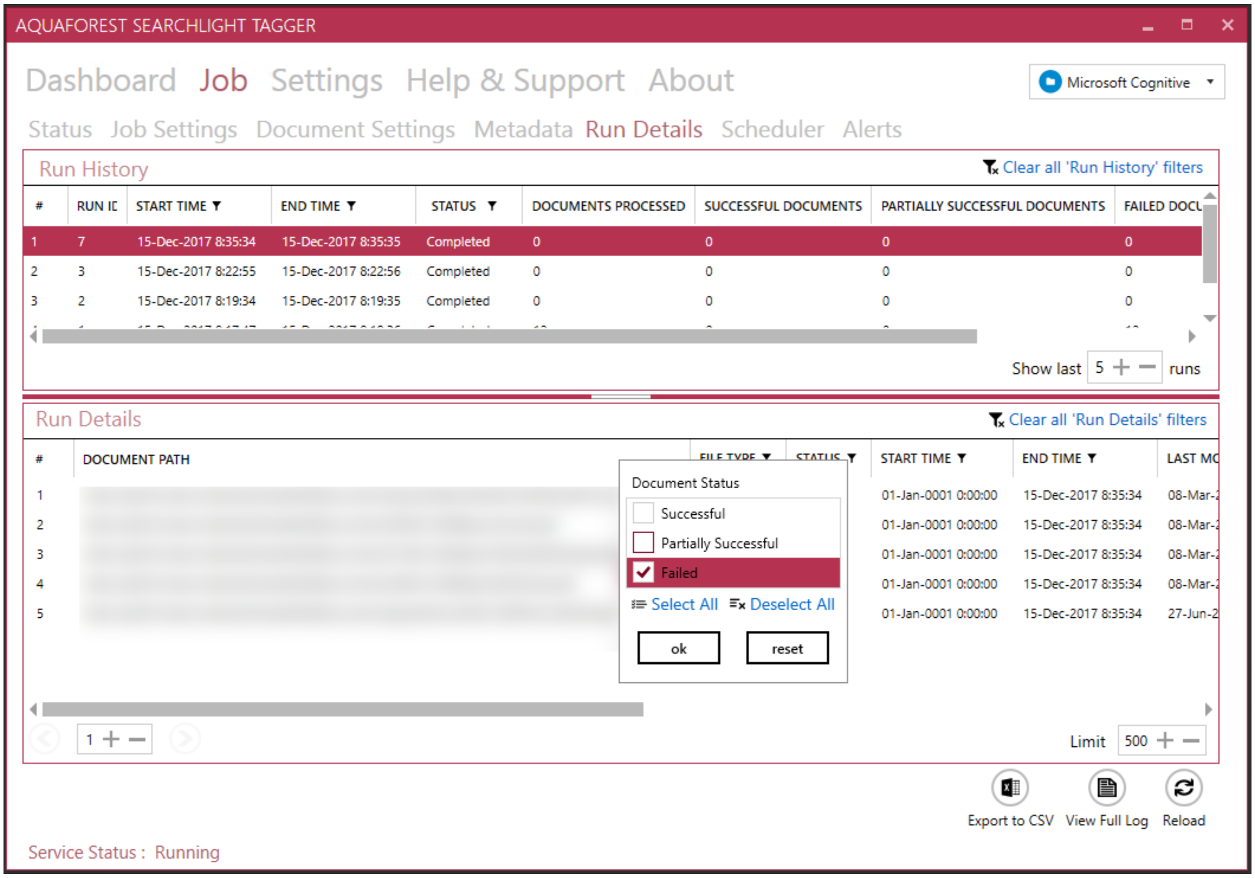
Show next 5 scheduled times

Friday, February 23, 2018 16:00 PM
Friday, February 23, 2018 20:00 PM
Saturday, February 24, 2018 00:00 AM
Saturday, February 24, 2018 04:00 AM
Saturday, February 24, 2018 08:00 AM

Service Status : Running

Once you are done, your job is ready to be run.

Of course, Searchlight Tagger has advanced logging features too, to support your health-check, debugging and troubleshooting needs.



#	RUN ID	START TIME	END TIME	STATUS	DOCUMENTS PROCESSED	SUCCESSFUL DOCUMENTS	PARTIALLY SUCCESSFUL DOCUMENTS	FAILED DOCU
1	7	15-Dec-2017 8:35:34	15-Dec-2017 8:35:35	Completed	0	0	0	0
2	3	15-Dec-2017 8:22:55	15-Dec-2017 8:22:56	Completed	0	0	0	0
3	2	15-Dec-2017 8:19:34	15-Dec-2017 8:19:35	Completed	0	0	0	0

Show last 5 runs

#	DOCUMENT PATH	FILE TYPE	STATUS	START TIME	END TIME	LAST MO
1	[redacted]			01-Jan-0001 0:00:00	15-Dec-2017 8:35:34	08-Mar-2
2	[redacted]			01-Jan-0001 0:00:00	15-Dec-2017 8:35:34	08-Mar-2
3	[redacted]			01-Jan-0001 0:00:00	15-Dec-2017 8:35:34	08-Mar-2
4	[redacted]			01-Jan-0001 0:00:00	15-Dec-2017 8:35:34	08-Mar-2
5	[redacted]			01-Jan-0001 0:00:00	15-Dec-2017 8:35:34	27-Jun-2

Document Status
- □ Successful
- □ Partially Successful
- ☑ Failed

Service Status : Running

Summary

With this white paper, my goal was to show you what options and supporting features you have in SharePoint to manage metadata fields, and how to tag and auto-tag documents. I hope that by now, you can see it's a complex topic, and providing good quality metadata is the common interest of everyone.

Aquaforest Searchlight Tagger might be of your help if you're in need of a lightweight solution to support your organization's needs by rich auto-tagging features.

About Agnes Molnar

Agnes Molnar, the managing consultant, and CEO of Search Explained, specializes in Information Architecture and Enterprise Search.

Agnes studied Information Technology at the *Budapest University of Technology and Economics.* During her remarkable career, Agnes has collaborated with numerous renowned companies on both national and international scale, architecting and executing dozens of Enterprise Search implementations for both commercial and government organizations.

Due to her passion and zest for the IT field, since the year 2008, Agnes has been consistently receiving the prestigious **MVP award (Microsoft Most Valuable Professional)**.

Agnes's basic aim is to create awareness among people regarding **search options, techniques, and best practices**. According to her, masses should know how to search for something they are looking for in a *short and precise time frame*. Also, the searched material should be *accurate and useful* for the respective individual. This she accomplishes through the blog of Search Explained by sharing all new and upcoming search tips and techniques, and related Information Management methods. Moreover, she delivers all this in a remarkably easy-to-understand way so that everyone can comprehend and relate to it.

Apart from her blog work, Agnes has also written as well as co-authored various **books** on SharePoint and Enterprise Search. (See more on her Amazon Author page.)

She has also delivered many **speeches** and **workshops** regarding Search and different Microsoft technologies around the globe, i.e. in Singapore, USA, Canada, UK, Denmark, Germany, Peru … and the list is endless. (See more on her list of Events.)

In her free time, she writes her personal blog at https://it-mompreneur.com , about the topics of entrepreneurship, mentoring, technical writing, public speaking, time management, and many more.