Strategies for SharePoint Metadata

01/2/2015

Brendan Clarke
The business benefits of adding metadata to your SharePoint content

Introduction

Analysts such as Gartner and IDC estimate that as much as 80% of the information within an organisation is contained in unstructured data such as documents and e-mails. As it has been traditionally hard to find, organise and analyse unstructured data it is often ignored in decision making. Any organisation who adds metadata to their content gains a significant competitive edge.

Much of this unstructured data is of very high value. Over 60% of organizations admit to storing a significant proportion of their electronic records in the native application format such as Word or Excel.

FIG 1. FILE TYPES FOR OFFICIAL RECORDS IN AN ENTERPRISE, * (C) AIIM.ORG 2011.
The benefits

The word metadata means “data about data” and for the purposes of this document this means additional terms (words or phrases) that are assigned to SharePoint documents in order to describe what they contain.

For an invoice document we might add three metadata fields: customer, year and country.

![Customer: Contoso](image)

![Year: 2014](image)

![County: UK](image)

FIG 2. AN INVOICE DOCUMENT WITH METADATA ADDED

Here are some of the areas that will be vastly improved with metadata added to your documents:

- **Search** (finding relevant documents)
- **Organisation** (displaying and sorting documents)
- **Compliance** (retaining or disposing of documents, eDiscovery and records management)
- **Insights** (Business Intelligence from documents)
Search

Introducing metadata allows the user to rapidly refine the search results to get to the documents they are looking for both quickly and accurately.

Using our invoice example we could allow the user to refine invoices by company names, dates or locations. In the figures below the users are presented with interactive charts allowing them to further refine their initial search for invoices.

FIG 3. A GRAPHICAL SEARCH REFINERS
Organisation

Most organisations have documents stored in folder structures. Saving documents in a folder structure introduces two major issues:

- People will save the documents to the wrong place or save it in several places
- Each folder only contains a subset of all documents making difficult to view documents according to different criteria (for example by year rather than by customer)

If we apply metadata to our documents we can quickly sort, filter and navigate on our metadata to show any documents we wish to see. We save all our invoices to a single location and use the metadata to view the documents in any way we wish. Here is how our list of invoices would look with some simple metadata applied:

![Image of invoice list with metadata applied]

**FIG 4. A LIST OF INVOICE DOCUMENTS WITH METADATA APPLIED**
SharePoint supports what is known as metadata navigation which allows you to refine and filter your view of all the documents in the library depending on what metadata values you are interested in. Here is how it would look if you wanted to know “What Contoso invoices were raised in the US in 2015?”

![Image of metadata navigation example with filters for company, financial year, and country applied.]

The user can select the company to refine by ① and then filter the documents on the financial year ② and country ③. When they select “Apply”, only the matching invoice(s) will be shown ④.
Compliance

If we are confident we can find information either by using search or metadata navigation then governance of documents (retaining, disposing and auditing of information) is vastly simplified as are the first steps of eDiscovery.

Information management policies can be applied to documents depending on the values of metadata. For example we could decide that all invoices raised in Germany must be retained for ten years, whereas the UK they should be retained for five years.

With metadata applied records management can be significant automated with records be stored in the correct locations with the correct governance supplied.

Insights

Introducing metadata to SharePoint documents in effects makes them structured which allows us to understand relationships and trends across large document sets. We can query and display information from related documents as if they were records in a database. Using our invoice example we can answer questions such as:

- Are we getting increasing numbers of invoices each month?
- How does the number of invoices in the UK compare to Germany?
- What was the average value of an invoice in 2014?

We are also able to create topic specific dashboards that relate to a given metadata value. For example we could create search driven pages that show:

- Sales performance for a region
- Sales of a given product per region
- Comparison of sales from two different date ranges or regions
Because the information is driven from the content of the documents the dashboards will always be up to date and require no IT work to maintain. Users can easily interact with this type of dashboard (sometimes known as a search driven application) to further drill into values and refine data.

**Conclusion**

The vast majority of information in your organisation is likely to be unstructured. Adding metadata will help you find, organise and gain insights from your documents. In the next section we will look at a number of methods for adding this valuable metadata to your SharePoint content.
Methods for adding metadata to your SharePoint documents

Introduction

This section will discuss the considerations for an approach and solution to applying this valuable metadata to your documents. One common area for all of the solutions we will cover is the requirement for SharePoint to store lists of metadata terms that will be used for tagging. This is known as a taxonomy (sometimes also referred to as the managed metadata service in SharePoint).

Taxonomies

A taxonomy (from the Greek phase meaning “Arrangement method”) is an area where an organisation can define lists of related terms to aid with classifying content. In SharePoint this area is known as the term store. A SharePoint taxonomy is organised into groups, term sets and the terms themselves.

![Taxonomy Diagram]

- **Groups**: top level container for groups of terms or term sets
- **Term sets**: a container for holding terms
- **Terms**: individual elements in a term set
The SharePoint term store is a central area for the creation and management of term sets, it was introduced in SharePoint 2010 and is also available in SharePoint online (Office 365). It supports importing existing term sets and allows administrator to delegate the management of areas of the taxonomy to other users.

Another key feature in the SharePoint term store is support for synonyms, this allows for a single term to have a number of derivations that will resolve to the same term. For example if we had a term set of company names we could have an entry for “British Broadcasting Corporation” with the synonyms “BBC”, and “B.B.C”. There is also full support for multiple languages when defining terms in the SharePoint term store.

**Approach considerations**

Before assessing a method for adding metadata to your content there are a number of very useful questions to be answered around the general approach:

- **Do we want to add metadata to everything?** Some documents, libraries or sites may be of higher importance when it comes to adding metadata. Picking only high value data or adopting a phased approach is good practice for any information architecture project
- **Do we know what we have?** An understanding of what content exists already and if any taxonomies have already been created within the organisation
- **Multiple languages.** Is a multi-lingual approach required for metadata tagging?
- **In house skills.** Do we have a data librarian, a SharePoint expert or in-house development skills
Solution considerations

There are a number of considerations when deciding the best method to apply metadata to your documents. We shall use the following criteria when comparing various methods:

- **End user burden** – how much extra burden (if any) does applying metadata to documents place on the users who are saving documents in SharePoint.
- **IT burden** – how much time will be required from IT both in preparation of a solution and ongoing maintenance
- **Solution cost** – if we are buying or building a solution what is the financial cost
- **Legacy considerations** – does the solution apply metadata to old data in SharePoint or just to documents that are added after the solution is implemented
- **Ability to scale or be complex** - can the solution be used to add a large amount of complex metadata to documents
Using the end user to apply metadata to your SharePoint documents

In this section we are going to look at the first method of applying the metadata – requiring the end user to manually add the metadata tags when they save their documents to a SharePoint library.

Introduction

This method is often the first technique an organisation will introduce when they decide they would like some simple metadata adding to their documents. The following steps are taken by IT in conjunction with the business:

(1) IT or the business define some term sets in the SharePoint term store

(2) In some or all of the SharePoint document libraries site columns are added into the library (of the type managed metadata) which link to the term store values

(3) Some fields are typically made mandatory (that is, the file cannot be saved without selecting a value)

Once implemented, when the end user saves a file to a SharePoint document library they will be presented with a number of fields to complete.
An organisation can choose which fields are optional and which are mandatory in order for the document to be saved, in the above example the customer name is a mandatory metadata field (marked with an asterisk).

The end user can enter the values manually (and SharePoint will offer auto-complete suggestions as they type) or they can click the tag icon for each and they will be able to choose from the list of pre-defined terms in the associated term set.
Assessment of the end user method

End user burden - High

The main downside of using the end users to add metadata documents is that it adds a time to the process of users saving their files. Even with a few simple fields to complete (often from drop-down menus) end users will often resent having to add metadata in order to save their document to SharePoint.

Best practice if you are asking users to complete metadata is to have as few fields as is possible for them to complete (ideally three or less). Wherever possible metadata should be automatically inferred from the library or content type. This is done by adding site columns to a library or content type and setting the default values. If you add too many fields for the user to complete, there are two very likely outcomes:

1. The user will stop saving documents to SharePoint completely preferring to save them locally and share them via e-mail. **A document management system is only useful if it contains documents**
The user will quickly work out the fastest way of completing the fields with the bare minimum of data (see below for the “asbestos problem”. **Bad metadata is worse than no metadata**

---

**The Asbestos problem**

“We recently worked with a consultant who was charged with introducing some simple metadata to project documents for a large construction company. In an attempt to keep things as simple as possible only one field was required to be completed to save the document, a field called “Project classification”. The consultant built a list of types of projects with 35 entries to choose from. The list looked like this:

Asbestos, Bridge, Churches, Damp proofing ... (and so on)

3 Months after introducing this plan they reviewed how the documents had been tagged and over 60% of them were tagged as “Asbestos”, the first choice on the list.

---

**IT Burden – Low / Medium**

IT will need to create and maintain the new taxonomy term sets and assign them to various document libraries. Assuming it is kept simple the overhead should not be too large but it becomes increasing onerous with large numbers of metadata tags.

**Cost - Low**

Aside from the cost of IT setting up the taxonomies and a productivity overhead of the end users there is no software to be purchased for this solution.
Legacy considerations - Poor

This solution works well for new document libraries but does not offer a solution for adding metadata to existing documents. The only real method is to edit each document and manually assign metadata. This is only practical for relatively small numbers of documents.

Ability to scale or be complex - Poor

End user tagging is only suitable for simple tagging of a small number of metadata terms.

SharePoint limitations

Assuming your taxonomies are created well and your end users are happy adding accurate and consistent metadata there are unfortunately a number of scenarios where SharePoint introduces a problem when it comes to end users being responsible for tagging. The following are some scenarios where metadata tagging does not happen when a user saves a file:

1. **Uploading more than one file at a time** – SharePoint allows users to drop multiple documents into a document library. If more than one document is added in a single upload then metadata requirements are ignored even if some fields are mandatory.

2. **File explorer** – If the user uploads files by mapping drives or using the file explorer then metadata is ignored.

3. **Microsoft Office 200x** – The support for adding metadata using Office client programs was introduced in Office 2010. Older versions of office will not be able to add metadata.
Summary

Implementing end user tagging of metadata is quick and has no cost. Additionally, the end user is often the right person to known the correct metadata to apply. The challenge is that end users do not like adding metadata themselves and will often avoid the process or add unreliable data. This method can work well for limited documents and a very small number of fields. There are real dangers however that users will stop using SharePoint to store documents.
Metadata tagging using SharePoint out of the box functionality

SharePoint 2013 offers some capability for automatic metadata tagging without the need for additional software. This is known as custom entity extraction and is in fact an extension of the search engine processing pipeline and the tagging happens whilst the content is being indexed. The metadata can be added to the search index as a managed property (a managed property is a field that can be used to extend the capabilities of SharePoint search).

The process for implementing this is done by IT as and is as follows:

- Create a list of matching entities (known as an entity extraction dictionary)
- Deploy the dictionary using PowerShell (this is not currently possible via the UI)
- Map the entity to a managed property

Full details of this process are here.

When the search crawler indexes the new document it will look for matches in the entity dictionary (there are four available types of matches). If matches are found then the associated managed property will have the metadata terms added.

Some limitations

Although this can be a very useful tool for simple tagging there are two limitations to be aware of:
1. **Not currently available for SharePoint online** ([details](#)). This option is only available for on-premise SharePoint deployments.

2. **The metadata is applied to the search index, not to the document itself.** If you are only applying metadata in order to improve search this may be sufficient but it would not allow the metadata to be used for metadata navigation and other areas of SharePoint.

### Assessment of the end user method

**End user burden - LOW**

The end user is not involved in the tagging process at all. When the search crawler indexes the document the metadata will be applied.

**IT Burden - MEDIUM**

IT will need to create and maintain the dictionary files. They will also need to deploy each file and map the managed properties for each. Will need to be able to deploy files using PowerShell.

**Cost - LOW**

This option is available out of the box with SharePoint 2013 on premises.

**Legacy considerations – GOOD**

The organisation can invoke a full search crawl which will examine every document and apply the metadata if there are matches in the dictionary file.

**Ability to scale or be complex – POOR**
In common with the end user tagging method using entity extraction can work well with a few simple metadata term sets. No logic can be applied to the tagging process, it simply looks for word matches. Maintaining a large number of dictionary files becomes an issue.

**Summary**

The fact that this method is not available with SharePoint online and that the metadata added to the search index rather than the documents themselves may make this unsuitable for many organizations (it is rarely used). For simple or small sets of terms it may however be a quick and effective solution for improving search.
SharePoint metadata tagging using rules based taggers

Previously we explored tagging which is done by end users or by SharePoint entity extraction and will now explore some of the software options for metadata tagging in SharePoint. Most of software solutions currently on the market are rules based and follow a process such as:

1. The organisation creates a number of taxonomies (for example, products, locations, departments)
2. For each term a rule or number of rules are defined to describe when a term should be applied

Types of rules

Broadly speaking there are generally three types of rules:

**Simple rules** – a simple tagging application may just look at terms in the SharePoint term set and look for matches in the document. This is similar to the SharePoint entity extraction discussed in the previous article but has two distinct advantages:

1. The terms are managed in the SharePoint Term store and not in text files – this makes working with the terms much more straightforward
2. The terms can be added to the documents themselves and not just a property in the search index
FQL based rules – **Fast Query Language** (FQL) allows simple logic to be applied to if a term should match or not. An organisation can define a rule for each term that has a number of elements such as other words should or should not exist in the same document. If this is carefully done the tagging can work very well. For example, if we wish to tag all instances of the word “Apple” as a computer manufacturer we could define the match as:

*Apple and (computer) and not (tree)*

This would tag a document with the tag “Apple” from the “Computer” term set only if the document contained both the words “apple” and “computer” and did not contain the word “tree”.

**Advanced rules** – Other software solutions provide methods for building advanced rules such as assigning a score or weight to multiple terms or creating complex logic and interdependencies between terms. When implemented correctly they can be highly effective in particular when complex taxonomies are required and a deep understanding of when a term should be applied is already understood.

Some solutions may also allow for pattern matching (for example to match reference numbers of any other entity with a fixed format). One of the most powerful ways of pattern matching is to use **regular expressions**. Although quite difficult to learn initially, regular expressions can be very powerful when implemented.
Assessment of rules based tagging

End user burden - LOW

The end user is not involved in the tagging process at all. The tags may be added in real time (when the user upload the document) or run as a SharePoint timer job or as part of the search indexing.

IT Burden - HIGH

Taxonomies will be need to be created and maintained. Additionally a rule for each term in a taxonomy needs to be created and maintained. This is usually an iterative process with ongoing tuning to get the correct metadata tags applied. Implementing rules based solutions typically requires a significant investment to prepare the pre-requisite terms and rules.

Cost – LOW/MEDIUM/HIGH

Solution costs typically vary depending on the complexity of the solution. More complex implementations will usually require a services / consulting element.

Legacy considerations – GOOD

All documents can be tagged with this method.

Ability to scale or be complex – GOOD

Rules based solutions can often manage a large number of term sets and can manage complex combinations of logic.
Summary

There are a number of vendors offering solutions for the automated application of metadata tags and they vary in complexity. Most require the taxonomies to be already created (some may have tools that will suggest common terms or phases in a given collection of documents) and may require rules to be created for each term. The project time and costs for creating complex taxonomies and rule sets can be high and will require ongoing maintenance and tuning. The pay-off however can be accurate tagging of content in many situations. Rules based methods work well in environments where strict document categorization is key to the business and where structure adds great value. If existing taxonomies and pre-defined rules already exist this may be a good option.
Semantic tagging

In this section we will look at a relatively new method for adding metadata to documents, the use of semantic based tagging. Semantic based taggers understand the meaning of words and phrases and do not require pre-defined taxonomies or rulesets. There are a number techniques employed for semantic based tagging such as natural language processing and machine learning. Semantic tagging engines have been trained on a vast number of documents (either automatically or with human intervention) and can produce very accurate metadata tagging.

Semantic tagging engines may offer a number of language processing functions such as:

- **Entity extraction** – The automatic identification of known entities within a document, examples of entities may be people or company names, locations and technical terms. These terms can be identified and tagged as metadata without the need for pre-defined taxonomies
- **Summarization** – The production of an accurate summary (usually a paragraph or less) of a longer document
- **Sentiment analysis** – The ability to discern and score the sentiment of a document or individual entity
- **Concept extraction** – Understanding of the concepts discussed in a document. For example a document may be tagged with the concept “Merger” despite that actual word not being present within the document
- **Multi-lingual** – some semantic taggers are able to automatically detect the language a document is written in. Even more advanced taggers are able to apply metadata tags in several languages
- **Word sense disambiguation** – A key difference between semantic methods and any other method of metadata tagging is that the tagger takes the context of each work into account when extracting entities. For example, given the sentence “You need to replace the batteries in your mouse”, it will be understood the mouse in question is a computer peripheral rather than a rodent.
Assessment of rules based tagging

End user burden - LOW

The end user is not involved in the tagging process at all.

IT Burden – LOW

Semantic tagging does not require taxonomies or rule sets to be configured before use, in fact some semantic taggers can create taxonomies from the content they process.

Cost – LOW/MEDIUM/HIGH

Solution costs typically vary depending on the complexity of the solution.

Legacy considerations – GOOD

All documents can be tagged with this method.

Ability to scale or be complex – GOOD

Semantic tagging solutions are highly effective at processing very large volumes of documents.

Summary

With the emergence of semantic based tagging many organizations will be able to achieve accurate metadata tagging without investing significant time and effort in preparing taxonomies and rules for tagging. Semantic taggers can discover relevant metadata tags as they process documents and build bespoke taxonomies on the fly. Applying accurate metadata without any burden on end users or IT departments is highly desirable in many organisations.
Other features such as summarization, sentiment analysis and language detection can add further valuable metadata to documents that is not available in traditional systems. Machine generated metadata is highly consistent which allows the data to be used in business intelligence scenarios.
The Termset platform

The Termset platform includes a tagging engine that has full semantic capabilities as follows:

- **Entity extraction** - hundreds of entities can automatically be tagged (all current entities are listed [here](#)) without the need for pre-defined taxonomies or rules

- **Sentiment Analysis** - Detailed and accurate scoring of the positive and negative sentiments for SharePoint content

- **Summarization** - Concise summaries created of documents.

- **Language Support** – Able to apply tags in 9 languages and can recognize over 90 languages (language details are available [here](#))

It additionally includes other methods of tagging:

- **Pattern Matching** – you are able to train the platform to recognize patterns in your text such as customer reference numbers, parts numbers or extract numeric values from documents. The platform ships with over 50 useful patterns (for example: postcodes, vehicle registration numbers and credit card numbers) and you can add as many others as you wish.

- **Term store terms** – If you have existing taxonomies defined you can import them into SharePoint and the tagger can utilize these terms when documents are processed. There is full support for multi-level term sets and synonyms.
Additional features

The Termset platform has been built from the ground up for SharePoint. It adds the following functionality directly into the platform:

- **Automated information architecture** – The platform will make recommendations for site columns, document library columns and most importantly metadata terms to be added across your SharePoint sites. With one click they can be created and maintained for you.
- **Taxonomy creation** - The platform will create and maintain SharePoint taxonomies unique to your content.
- **Rich data visualization tools** – The platform allows users to create and share rich interactive views of documents (known as a lens, more details are [here](#)).
- **Cloud based** - No software to install or additional servers required. Free trials of the platform are available [here](#).

For further details of the Termset platform or to request a free trial please visit [www.termset.com](http://www.termset.com)