

White Paper



Enhancing Search in SharePoint 2010 Using Managed Metadata Services and Analytics

Prashanth Govindaiah, Abirami Rajendran

Abstract

Metadata is additional data/information about other data/artifact which is used to describe it better. However this data resides outside of the artifact itself. E.g. Author, department to which a document is written by or belongs to, type of the document like white paper/battle card, etc. are the additional data about the document itself but resides outside of the document and speaks more about the document.

Metadata¹ Management is one of the critical aspects of any organization which has a sizeable chunk of content management with it. By managing its metadata, consistency of the information used across the various units and teams in an organization could be enhanced, thus facilitating the information to be more locatable.

This white paper describes the challenges of how metadata can be managed and also includes real-world techniques for instigating an enterprise metadata approach using SharePoint Server 2010. Managed Metadata Services will be a giant step for the management of metadata as significant search and filters are enabled in SharePoint 2010. Challenges described here uses Microsoft Office SharePoint Server 2007 as reference Content Management Product though the argument holds good for most of the other content management products as well.

For more information, Contact askus@infosys.com

Scenario

For better understanding of the challenges and solutions let us consider the following scenario of Infosys as an organization. Infosys has 2 departments namely Banking and Capital Markets (BCM) and Finacle and let us try to understand the challenges that might arise and how they are solved by leveraging the features of SharePoint 2010.

Setbacks in the existing systems

The main rationale for defining metadata is for enhancing the storage and retrieval of an organization's most precious resource — information. A comprehensive evaluation of the organization's information architecture can help identify potential inefficiencies, such as the following:

1. **Lack of standardized metadata across the organizational sites, applications:**
This leads to inconsistent way of tagging data. Thus, poorly catalogued and managed storage of data can cause end users or developers to locate and bank on the incorrect data or assume that the data is missing. For example, in an organization to fill in the field called Unit, one person might fill in as BCM, second as Banking, third person as Capital Markets. When somebody tries to get all documents belonging to this unit BCM, he may not get the right set of documents.
2. **Storing metadata hierarchically:** Many content management systems either do not support hierarchical metadata or they are not used by the organizations. For example, If one has to tag a document stating that it belongs to Finacle project, Banking Unit of Infosys so that he gets back the document in result set irrespective of which field I enter in the hierarchy, it is not possible today.
3. **Free text Metadata:** Though the metadata should be standardized and governed, also there should be scope provided for end users to add metadata tags so that search relevance is improved as their inputs also get crawled and indexed which is missing today. Inconsistent use of metadata such as free text without proper governance can make it strenuous to search for and compare related data or content.
4. **Grouping Metadata:** This is about managing Metadata differently for different sub units in a big organization where each sub unit almost acts as a different company by itself. Considering a real time example, in an organization, the Sales department is connected to the company's Global store as well as to their own Sales Department's Managed Metadata store. So, while searching for a particular term, all files associated with that term are retrieved from the multiple site collections across the web applications.
5. **Synonymous Terms:** People often use terms synonymously for searching, for example, WSS, WSS 3.0, Windows SharePoint Services, Windows SharePoint Services 3.0, etc. Although different search keywords are used to refer to the same document or file, the metadata might be defined as one of these terms and the end result might be that it may not be returned as a search result.
6. **Lack of Search Administration Reports:** Administrators experience difficulty in monitoring the search service applications as there was no specific monitoring feature to oversee the health of search service applications on a SharePoint farm. Also, there are no provisions for an administrator to actually view a graphical representation of the searched keywords against the time taken for its retrieval. These posed a challenge for the administrators to enable search enhancement without high-level and in-depth monitoring of data.

The key limitation of MOSS 2007 was the site collection boundary. Site columns were specific to a site collection. If one wanted to share metadata and taxonomy² across multiple collections, they had to necessarily replicate the same lists in individual collection. This issue would be compounded when updates need to be done for metadata across multiple collections.

Also MOSS 2007 was incapable of addressing the above mentioned challenges namely no hierarchical metadata, no sharing controlled vocabularies across site collections, etc.

In reality, it is essential that the Enterprises need to have a bird's eye view of their data across all of their repositories irrespective of its location either on the user's PC, a shared server or SharePoint.

To overcome such challenges; SharePoint 2010 has been bundled with a new important feature called **Enterprise Metadata Management (EMM)**. EMM is a key building block in Enterprise Content Management and will offer enterprise, management of metadata terms and content types for the content management systems. This will permit the sharing of the metadata taxonomies and terms in organizations across its multiple SharePoint websites² and site collections³.

In addition, content types⁴ are allowed to be shared across websites and site collections using the uniform service. This will provide a much more amalgamated view of data across the organization and thus necessitating the requirement for managing the metadata across the enterprise. Also, there is a new feature that allows management of metadata on the file share with the new File Classification Infrastructure (FCI) ⁷ in Windows Server 2008. As documents are passed around the enterprise from the user's PC to SharePoint, a common metadata tagging approach is carried out.

It would come into view that SharePoint 2010 embraces many new features under the [Enterprise Metadata Management](#), and appears to have addressed some of the above key challenges.

Metadata in SharePoint 2010 is broadly classified into two types.

Figure 1: Classification of Metadata



Governance Driven (Taxonomy): These terms are usually pre-defined by an enterprise administrator and are centrally governed. In other words, these terms are [Managed Terms](#). A controlled term can only be created by those with appropriate permissions. Managed terms can be structured into a hierarchy and can be selected by the users out of the new 'Managed Metadata' column type. When a metadata column of this type is defined, the user will be permitted to decide the metadata from the collection of pre-defined managed terms. *Term sets*⁶ (think of them as taxonomy facets) are a compilation of associated terms that can be hierarchically structured.

User Driven (Folksonomy): These are words or phrases that have been added by users to SharePoint 2010 items. In other words, these terms are [Managed Keywords](#). Keywords are often used in more ad-hoc folksonomies where it is permissible for the user to tag an item with any keywords they deem are appropriate. Keywords are not represented in a hierarchy. The Managed keyword column is present by default in a number of SharePoint 2010 content types which implies that the administrators need not add it explicitly. User-generated keywords (aka tags) are kept in a non-hierarchical list known as the *keyword set*.

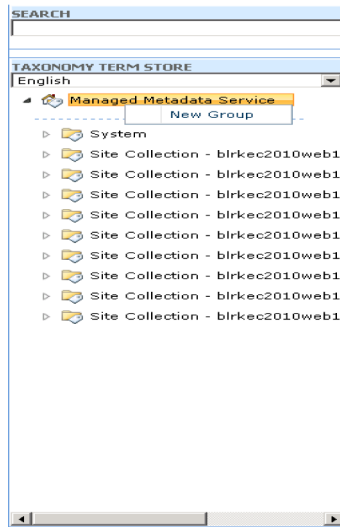
Solution for the Infosys scenario:

There are three managed metadata services created, one comprehensive for the entire organization, one exclusive to the BCM and one exclusive to the Finacle. Each unit has its unique SharePoint web with multiple site collections. The Units have connected their web application to several metadata services subsequently using the global terms plus their own local terms.

Below is the exemplification of an organizational instance, where in a user needs to enter in a field called "Unit" and the possible keywords are already available as a centrally governed keyword thus enabling classification and search easier.

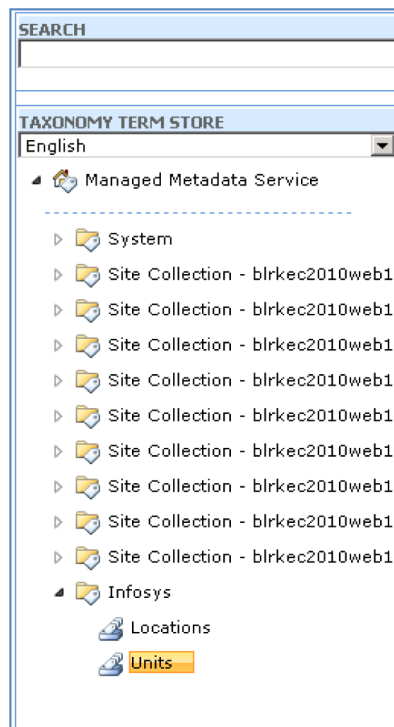
1. Create a new group (Infosys) in the Taxonomy term store under the Managed Metadata Service and create two Term Sets viz, Locations and Unit as shown in Fig. 3.
Thus, we here initiate the creation of a hierarchy level.

Figure 2: Step1: New Group Creation in Managed Metadata Service Application



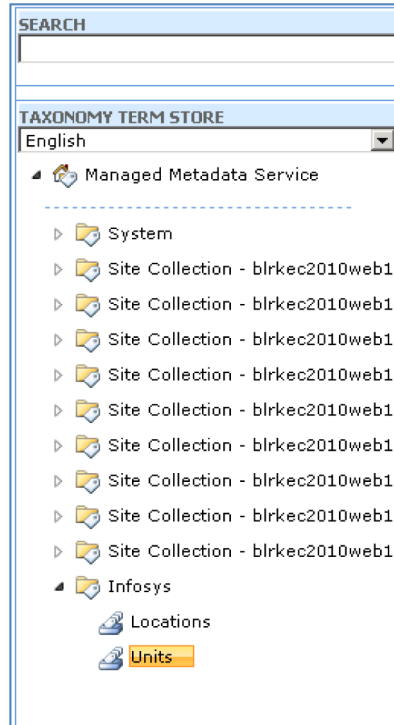
2. Create child terms for each term. By creating Child terms metadata across the organizational sites /applications is standardized.

Figure 3: Step2: New Term Set Creation under group



3. Create Synonyms⁸ for each term. Creating Synonyms for terms help to improve search relevance as their inputs also get crawled and indexed thus achieving greater scope for end users to retrieve the documents with ease.

Figure 4: Step3: New Terms Creation under Term Sets



4. After the synonyms (for many terms) are created, a custom list is now created in the site and the “Type of information in this column” is chosen as “Managed Metadata” and select to “Display the Entire path in the field” as the display value.

Figure 5: Step4: Defining additional properties like synonyms for new terms

The screenshot shows the 'PROPERTIES' form for a term named 'USA'. The form includes several sections:

- Available for Tagging:** A checkbox is checked, indicating the term is available for use by end users.
- Language:** A dropdown menu is set to 'English'.
- Description:** A text input field is empty.
- Default Label:** A text input field contains 'USA'.
- Other Labels:** A text input field contains 'America, United States, U.S.'.
- Member Of:** A table showing the term's membership in other term sets.

Term Set Name	Term Set Description	Parent Term	Source Term	Owner
Locations		Americas		ITLINFOSYS\Abirami_

A 'Save' button is located at the bottom right of the form.

5. Select the relevant managed term set to use in that respective column.

Figure 6 Step5: Defining Metadata column as a Managed Metadata column in a list

Unit

The type of information in this column is:

- Single line of text
- Multiple lines of text
- Choice (menu to choose from)
- Number (1, 1.0, 100)
- Currency (\$, ¥, €)
- Date and Time
- Lookup (information already on this site)
- Yes/No (check box)
- Person or Group
- Hyperlink or Picture
- Calculated (calculation based on other columns)
- External Data
- Managed Metadata

Description:

Require that this column contains information:

Yes No

Enforce unique values:

Yes No

Add to default view

Allow multiple values

Display Value:

Display term label in the field

Display the entire path to the term in the field

6. Now that there is a managed metadata site column, on tagging of a document from an appropriate library, the user can get a 'type-ahead' experience where suggestions would be obtained from the allowed terms. This showcases the standard functionality in SharePoint Server 2010, "Navigation by metadata" within a list or library – "navigation hierarchies" and "key filters" that prove to be highly beneficial when on a lookout for items in large lists.

Figure 7 Step6: Assigning the right node from Taxonomy to the Metadata column

Use a managed term set:
Find term sets that include the following terms.

Managed Metadata Service

- Infosys
 - Locations
 - Units
- Site Collection - blrkec2010web1-2009
- Site Collection - blrkec2010web1-2222
- Site Collection - blrkec2010web1-2345

7. Alternately the user can click the icon on the right and use a picker to select as shown in Fig. 8.

Figure 8: End User View of the Metadata column

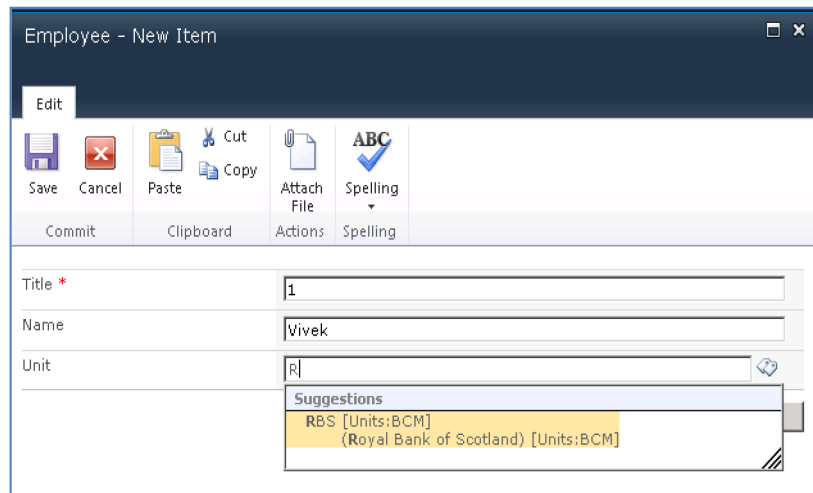
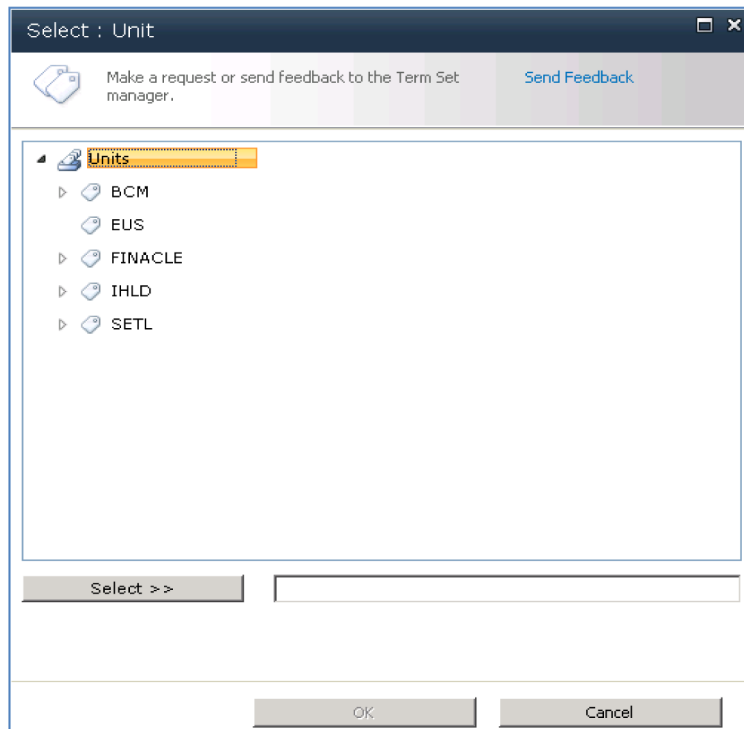


Figure 9: End User View (2nd View) of the Metadata column



Role of Analytics

It is a widely accepted fact that efficient metadata models improve search. To achieve this, it is required that not only metadata should be populated for all content but also a continuous monitoring of the search service applications should be done. This is achieved in SharePoint 2010 through

Search Administration Report

The Search Administration Report helps user to determine the health of search service applications on a SharePoint farm. Three different search administration reports viz., Basic, Advanced and Verbose search administration reports aid in structured monitoring.

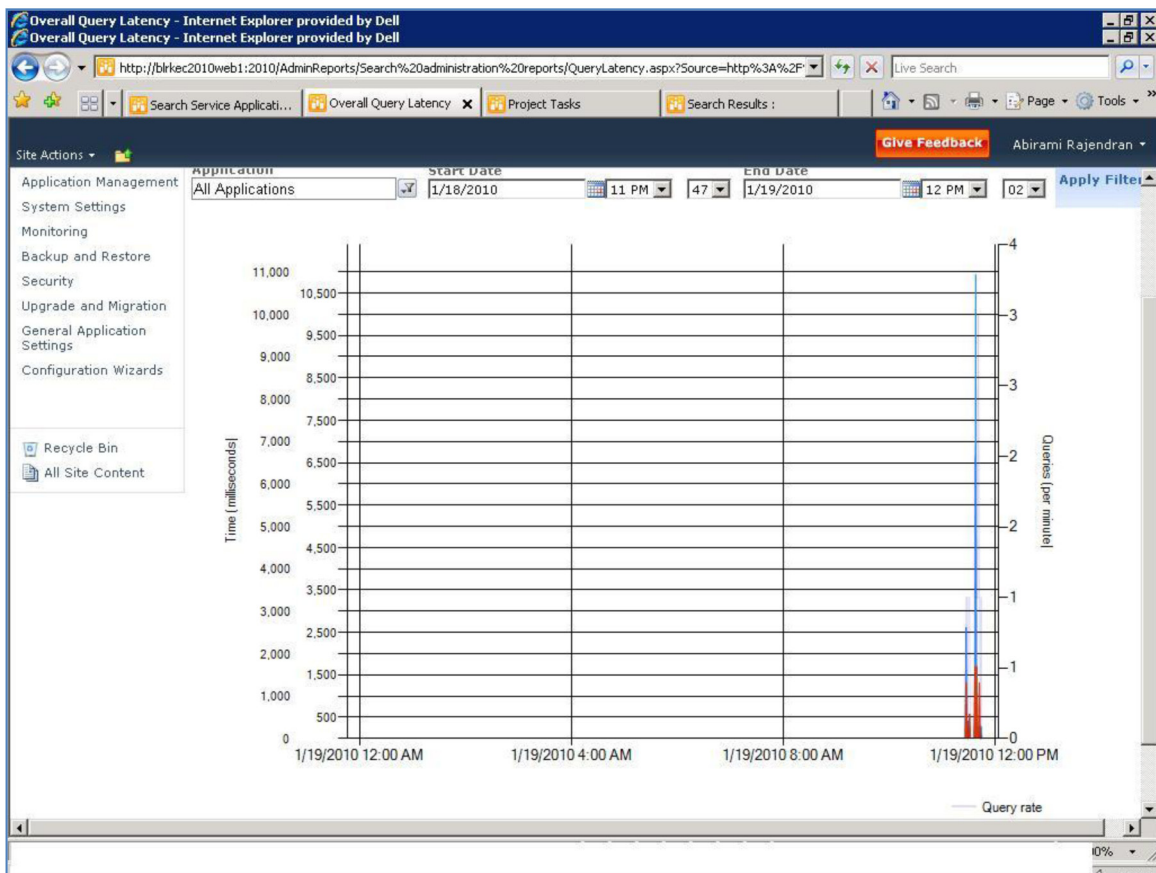
Scenario

A user needs to search an item from a predefined list; search for a term that is already present in the list and another term that is not present in the list. [The Basic Search administration report](#) provides a high-level monitoring data that is collected from all web apps that is created for the particular search service application. One of the features is, users can see a graph plotted for the query that is searched against the time that was taken to search and retrieve the query.

Obviously the time taken to search an item in the list is always lesser than to search an item that is not present in the list.

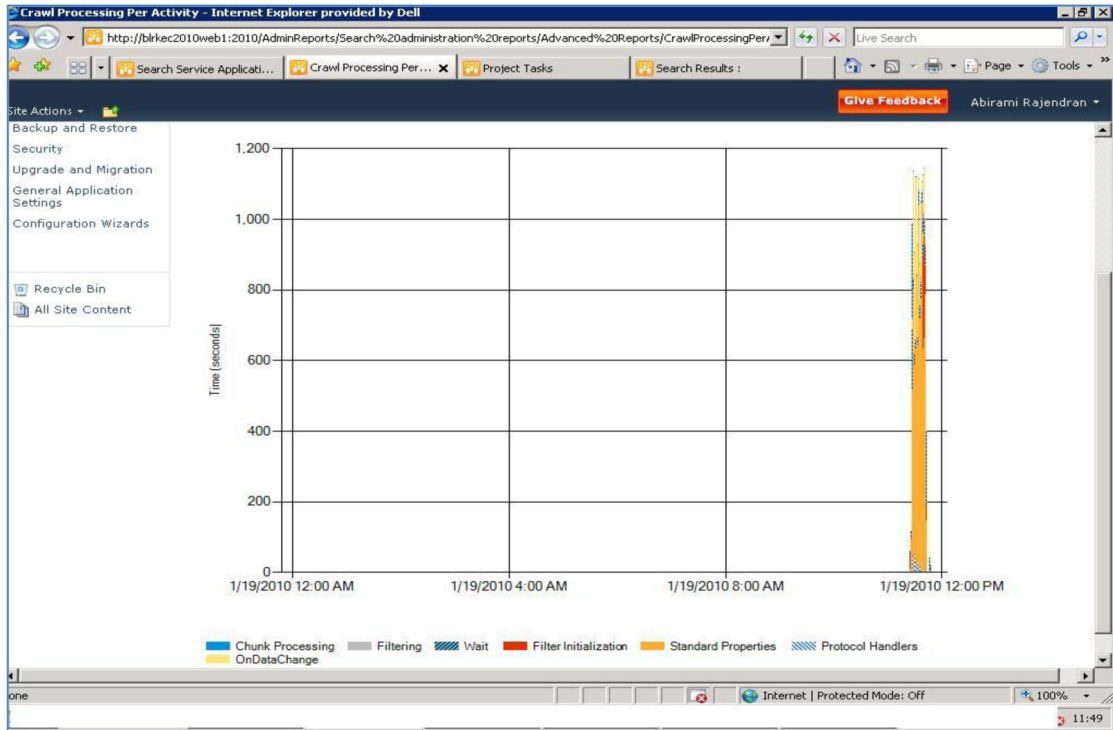
This graphical representation helps the administrator to understand the time complexity of the search, index an item thus giving an overview if the list needs to be broken down to enhance the search.

Figure 10: Overall Query Latency



Considering the same scenario where an item is searched from a pre-defined list, the [Advanced Search administration report](#) displays an in-depth monitoring data that is collected from all web apps that was created for the particular search service application. In this example, when the user searches for an item from a list, report displays the graphical plot on the time that is spent on filtering the word from the list, breaking down words in case of multiple word items, processing those words, etc. thus enabling the administrators know about the time taken for the search processing to occur for the particular term.

Figure 11: Crawl Processing Per Activity



The [Verbose Search administration report](#) is more similar to the [Basic Search administration report](#) and it also displays the percentile amount of time the search takes to execute the query and return a result set on the server-side in addition to the current crawl rate.

Thus, this helps in tailoring the system to meet the needs of the users, define how they use and ascertain information, and also to produce targeted content for your sites. Also Administrative reports, the pre-built reports, studies the usage data to examine the various facets of Microsoft SharePoint Server 2010, such as search crawl and query performance thus providing an insight into the search effectiveness.

Conclusion

The benefits of using managed metadata could be summarized as follows:

- Enables uniform usage of terminology as well as consistent usage of enterprise keywords that helps in better search relevance.
- Ensures dynamic interaction because using terms helps to keep SharePoint Server items cohesive with the business even as business keeps changing.
- Operates as the base for a colossal assortment of functionality transversely through the collaboration and publishing sites, and the whole lot in between thus enabling to implement end-user functionality in site. The outcome is not only to have the content tagged, but also effectively use it.

Glossary:

1. **Taxonomy** is a particular classification in which arrangement is in a hierarchical structure. Typically this is organized by super type-subtype relationships known as the generalization-specialization relationships, or more commonly as the parent-child relationships.
2. A **web application** is the logical & physical partition/container within IIS to create portal.
3. **Site collection** is a group of sites which can be managed together hierarchically. **Sites** that are present within a site collection contain common traits, like the shared permissions or galleries for templates or content types or Web Parts, apart from sharing a common navigation. A site collection comprises of a single top-level site, with any number of subsites organized hierarchically. A subsite is basically a distinct SharePoint site inside a site collection.
4. A **content type** is a reusable collection of settings that you would require to employ to a certain group of content. Content types facilitate in metadata management and performance of a document, item, or folder type in a reusable and centralized way that necessitate to be precisely the same across the site collection, like for example, a company logo or a brand name. For instance, if I type a document for the Banking Unit with some content and another document for the Insurance Unit, although the template of the documents, the workflows on it and the metadata tagged might be different, but the content types itself are related to one another.
5. A **“term”** is a basic construct — which means an expression that can be in association with content. A term can be a managed term or a managed keyword. For example, common things that user will want to sort or filter documents on. If users are likely to filter documents by the units that the document is associated with, then “BCMX” could be metadata.
6. **FCI** is an in-built result for classification of files that makes possible the manual processes for data classification to be automated with predefined policies centered on the significance of that particular data with respect to the on-going business. The Out of the Box Features of the FCI provides us the ability to define various Classification properties which integrates with SharePoint Server.
7. A **Synonym** is a term that can encompass any number of meanings. So if you fancy your documents being tagged as ‘SharePoint Workspace’ instead of ‘Groove’, you’d describe the latter as a synonym of the former. This ensures that the artifact will be a portion of the result set irrespective of the search criteria being either of the synonyms

About the Authors

Prashanth Govindaiah

Prashanth Govindaiah is a Senior Technology Architect who has been working on Microsoft technologies for most of his career. His roles in the last ~12 years at Infosys involved Architecting, Designing and developing software applications in various verticals. Work experience spanned across diverse organizational functions, organizational strategies, software service models, processes, systems, and structures. Last 8 years have been spent in designing and developing application systems, based on Microsoft development tools and technologies. Have exposure to pre-beta stage Microsoft’s development tools and their competitive products. He currently focuses on SharePoint 2010 building Migration and Governance tools for the same.

Abirami Rajendran

Abirami Rajendran is a Systems Engineer having 1.6 years of experience. She started working on MOSS 2007 and moved on to the Technology Adoption Program(TAP) of SharePoint 2010 and have been concentrating on SharePoint 2010 ever since. Currently, she focuses on building solutions for projects in SharePoint 2010. Apart from being a techno-geek, writing articles and stories, blogging and reading non-fictions are her hobbies.



For more information, contact askus@infosys.com

About Infosys

Many of the world's most successful organizations rely on Infosys to deliver measurable business value. Infosys provides business consulting, technology, engineering and outsourcing services to help clients in over 30 countries build tomorrow's enterprise.

For more information about Infosys (NASDAQ:INFY), visit www.infosys.com.