How to Use SharePoint Metadata to Improve Search and Control Content

by Mark Klinchin CTO, MetaVis Technologies

January 2010

MetaVis
TECHNOLOGIES

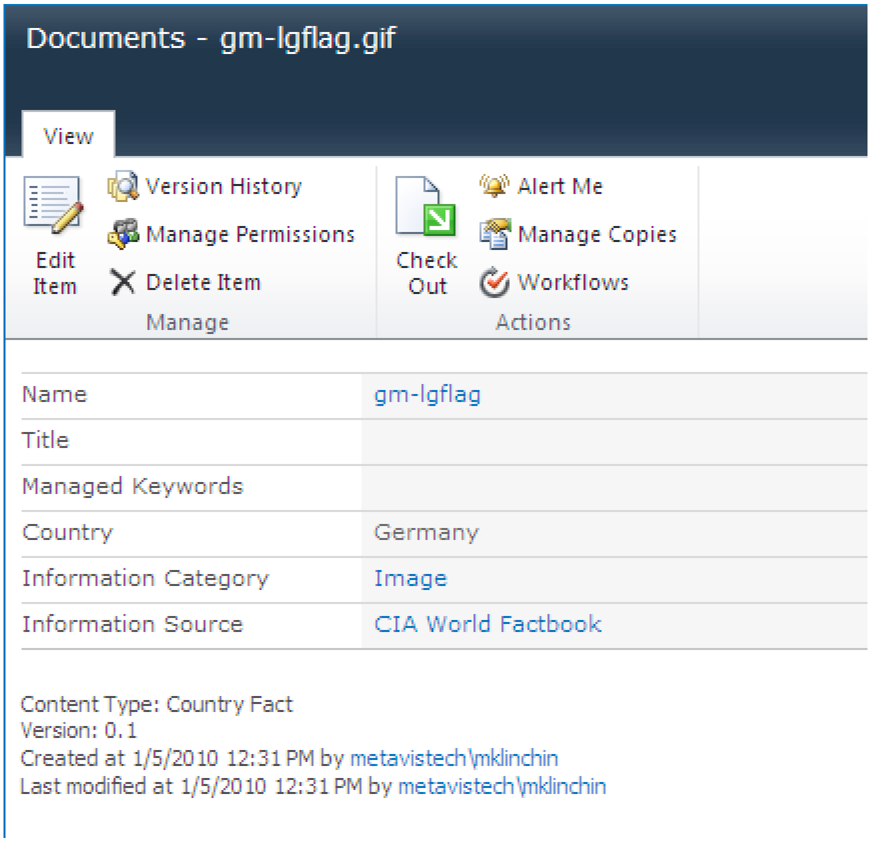
Phone: (610)-717-0413
Email: info@metavistech.com
Website: www.metavistech.com

● *Introduction to a SharePoint Metadata Model*

One of the main reasons for using a content management system like SharePoint is to efficiently find the right documents and to enforce well-defined processes that govern documents' behavior during different stages of their life cycle. Both of these activities would greatly benefit from the use of metadata, especially if the metadata model is properly designed and the values are consistently applied to documents.

Metadata typically contains several names/value pairs called *attributes* that describe a particular document. For example, an attribute name could be "Keywords" and the values could be specific keywords associated with certain documents like a web page or MS Word file. In more complex examples, metadata may have tens of attributes specifying different aspects of the document origin or use, such as Vendor, Country, Information Category, Source, Retention Policy, Archival Date, etc.

Figure 1 shows metadata for the document **gm-lgflag.gif** . This document has the following attributes: "Country" with the value *Germany;* "Information Category" with the value *Image*; and "Information Source" with the value *CIA World Factbook*.



Documents - gm-lgflag.gif

Name	gm-lgflag
Title	
Managed Keywords	
Country	Germany
Information Category	Image
Information Source	CIA World Factbook

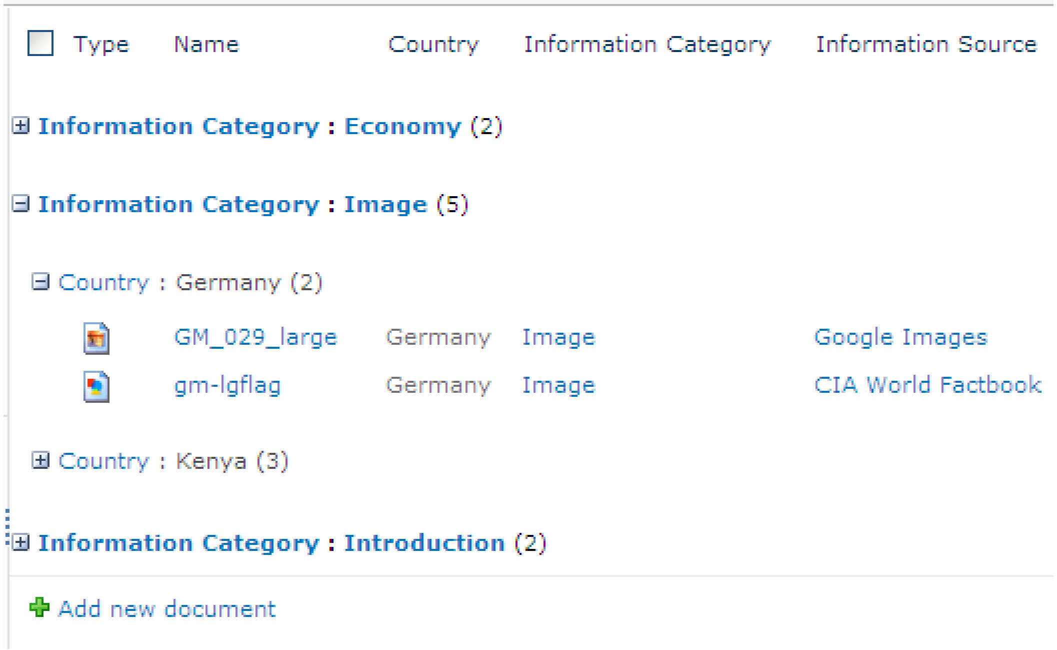
Content Type: Country Fact
Version: 0.1
Created at 1/5/2010 12:31 PM by metavistech\mklinchin
Last modified at 1/5/2010 12:31 PM by metavistech\mklinchin

Figure 1

Metadata models specify which attributes a certain type of a document may have and which attributes it must have. Metadata models are usually specific to a business case and vary between companies and even projects. As an example, we may define that all our web pages may have ‘Keywords’ as an attribute and all ‘Auto Part Contracts’ must have Vendor, Make, and Model and an optional Address and Phone Number.

Documents find-ability in the enterprise environment involves navigation through lists of folders and documents or filtered document sets; faceted navigation structures; or parametric and full text search results. There are two mechanisms that are the foundation of these techniques: extracting and indexing the usable text from the document (which may not be even possible in some cases i.e. pictures or videos) and using human manageable metadata to describe documents (that could be auto-populated using the text extracted from the document).

Figure 2 shows documents in SharePoint list grouped by *Information Category* and *Country* attributes.

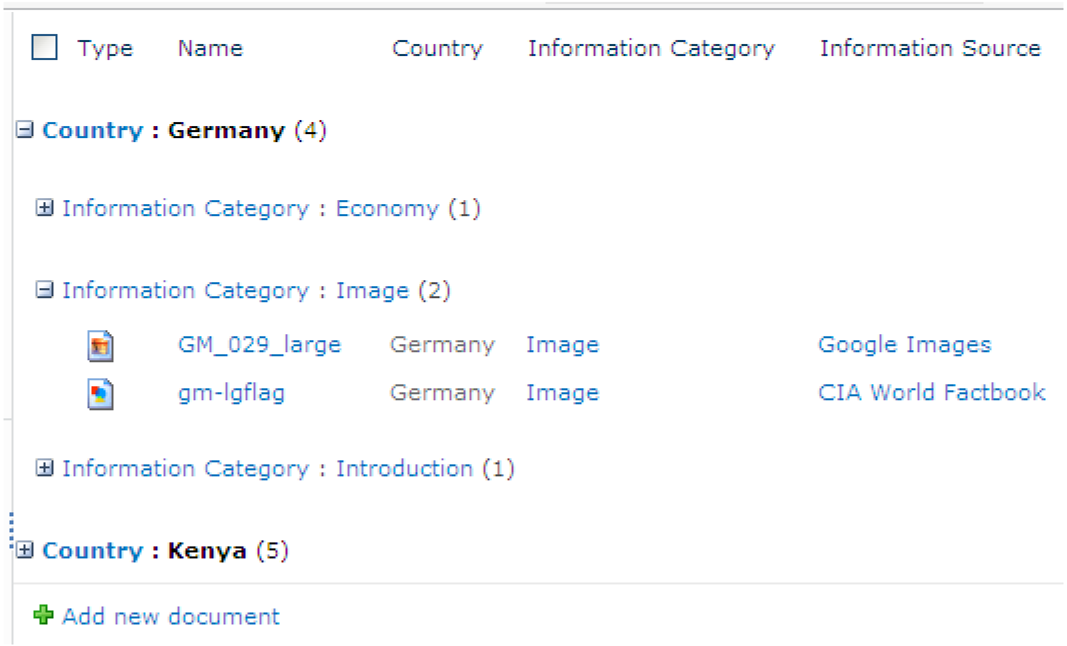


□ Type	Name	Country	Information Category	Information Source
⊞ **Information Category : Economy** (2)				
⊟ **Information Category : Image** (5)				
⊟ Country : Germany (2)				
	GM_029_large	Germany	Image	Google Images
	gm-lgflag	Germany	Image	CIA World Factbook
⊞ Country : Kenya (3)				
⊞ **Information Category : Introduction** (2)				

Add new document

Figure 2

Figure 3 shows the same documents as seen in Figure 2 in the same list but grouped by different criteria: *Country* first and then *Information Category*. These two views of the same content can be switched with a single mouse-click and provides different navigational experience for the user.



□ Type	Name	Country	Information Category	Information Source
⊟ Country : Germany (4)				
⊞ Information Category : Economy (1)				
⊟ Information Category : Image (2)				
	GM_029_large	Germany	Image	Google Images
	gm-lgflag	Germany	Image	CIA World Factbook
⊞ Information Category : Introduction (1)				
⊞ Country : Kenya (5)				

Figure 3

Content control functions include document workflows, security, printing, archival or retention policies, publishing and other mechanisms that enforce proper document capture, authoring, storage, distribution and disposal. To maintain content states and control information, control activities rely on information about the document, which is stored as a part of the metadata. Since control activities tend to vary from organization to organization, so does the metadata associated with them.

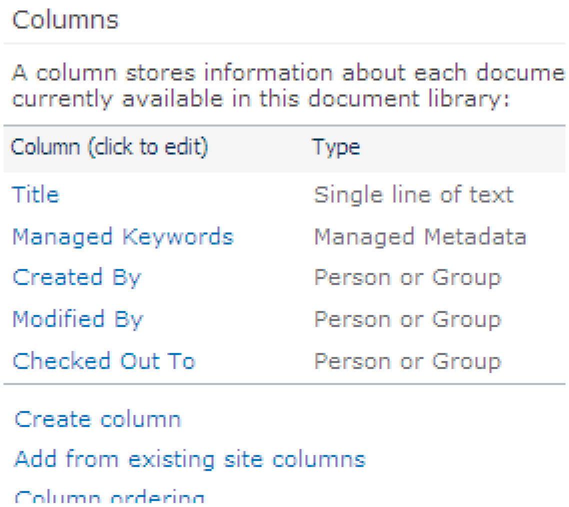
SharePoint implements metadata using columns, content types and lists. Efficient taxonomy design that leverages these objects (along with a proper site structure) targeted towards specific business tasks - will lead to a better use of existing content, increased productivity of information workers and a quicker return on your SharePoint investment.

Part II

● *Optimizing SharePoint Architecture to Improve Search and Control Content*

Each SharePoint site includes an out-of-the-box content type hierarchy and predefined columns. In the same fashion, each list comes with certain pre-defined content types and columns. Most of the out-of-the-box content types are generic. They define basic columns necessary to describe content. Examples include name and title columns for a document or date and description for calendar events.

Figure 4 Shows columns that are provided with an out-of-the-box SharePoint document library. It includes generic columns like *Title* or *Created By*.



Columns

A column stores information about each docume currently available in this document library:

Column (click to edit)	Type
Title	Single line of text
Managed Keywords	Managed Metadata
Created By	Person or Group
Modified By	Person or Group
Checked Out To	Person or Group

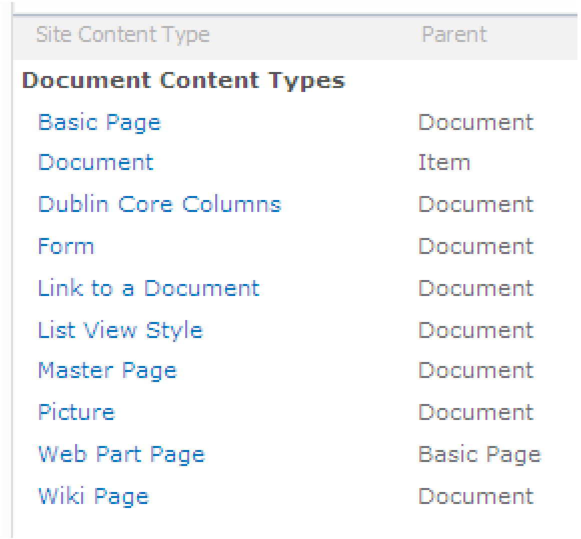
Create column

Add from existing site columns

Column ordering

Figure 4

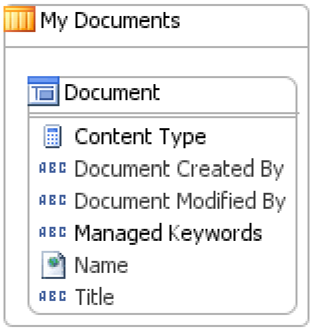
Figure 5 Shows SharePoint Content Types Gallery with out-of-the-box content types



Site Content Type	Parent
Document Content Types	
Basic Page	Document
Document	Item
Dublin Core Columns	Document
Form	Document
Link to a Document	Document
List View Style	Document
Master Page	Document
Picture	Document
Web Part Page	Basic Page
Wiki Page	Document

Figure 5

Figure 6 Shows a graphical depiction of document library *My Documents* before any customizations with the default content type *Document* which includes generic columns like *Title* or *Document Created By*.



My Documents

Document

Content Type
Document Created By
Document Modified By
Managed Keywords
Name
Title

Figure 6

Figure 7 Shows a graphical diagram of some of the out-of-the-box content types that are available in a default SharePoint site. It contains content types like Item, Event, Document, Folder and Basic Page with columns defined. The diagram also illustrates the parent-child relationships between SharePoint content types. Note that the Title column defined in the Item content type will be available in all subsequent content type due to SharePoint content type inheritance rules.

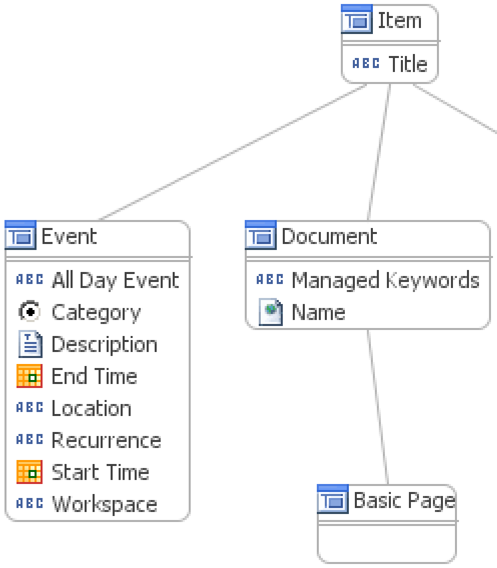
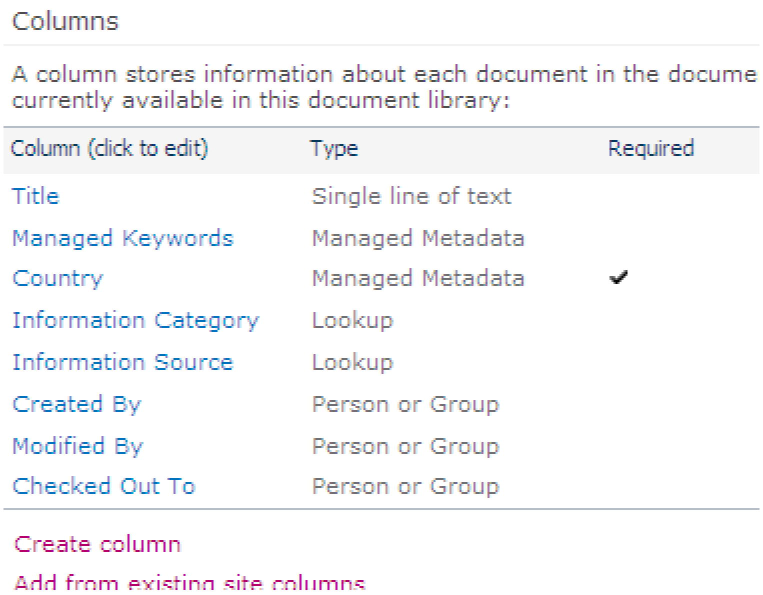


Figure 7

A common approach to enhance or customize the available set of fields would be to insert additional columns directly into lists and thus collect more information about the content stored there. Furthermore, adding these columns to list views will facilitate navigation through the content and enable the use of parametric search and/or faceted navigation in the list.

Figure 8 Shows list settings page for a document library after custom fields Country, Information Category and Information Source were added. Users may assign values for these fields for each item in the library.



Columns

A column stores information about each document in the docume
currently available in this document library:

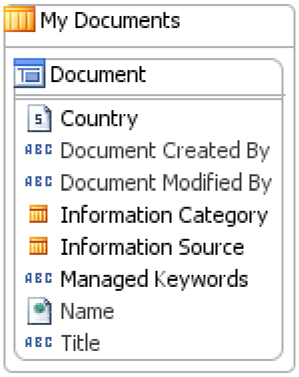
Column (click to edit)	Type	Required
Title	Single line of text	
Managed Keywords	Managed Metadata	
Country	Managed Metadata	✓
Information Category	Lookup	
Information Source	Lookup	
Created By	Person or Group	
Modified By	Person or Group	
Checked Out To	Person or Group	

Create column

Add from existing site columns

Figure 8

Figure 9 Shows a graphical diagram for the document library My Documents with the same custom fields described in the Figure 8.



My Documents

Document

Country

Document Created By

Document Modified By

Information Category

Information Source

Managed Keywords

Name

Title

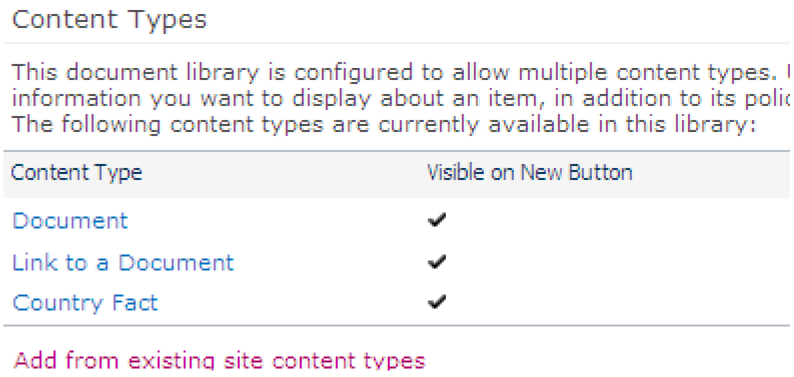
Figure 9

A common approach to organize SharePoint is to create a new site for every division, project, products, etc. In this case, the ideal tactic is to identify typical metadata sets, which can be implemented as content types on the site or site collection level, and reused in lists throughout the site hierarchy.

Identifying common metadata patterns that repeat over and over again allows an organization to develop a consistent approach to classifying and consequently "finding" their content throughout the SharePoint environment.

A similar methodology to list design will result in a standard approach to content accumulation and tagging. Lists with the similar columns should be redesigned in a way where these columns are aggregated in content types on a site level and then pushed down back to the lists. These changes will greatly improve faceted navigation, parametric and basic search for all items inherited from the base content type. Inherited content types also allow the use of generic workflows and policies applicable to all such items.

Figure 10 Shows document library settings page with three content types assigned to this library: *Document*, *Link to a Document* and *Country Fact*.



Content Types

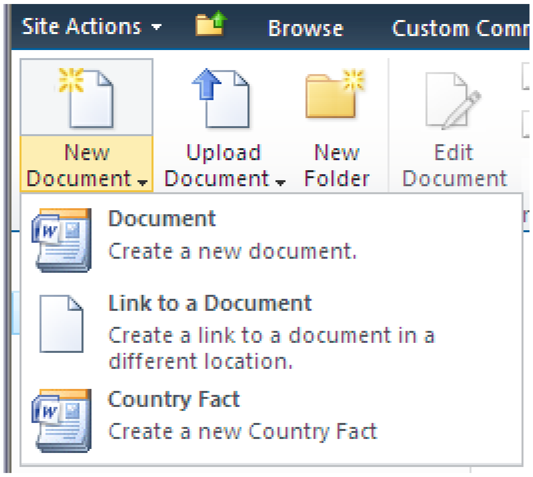
This document library is configured to allow multiple content types. I
information you want to display about an item, in addition to its polic
The following content types are currently available in this library:

Content Type	Visible on New Button
Document	✓
Link to a Document	✓
Country Fact	✓

Add from existing site content types

Figure 10

Figure 11 Shows how the Create *New Document* menu appears after document library is configured to include three content types shown in Figure 10



Document
Create a new document.

Link to a Document
Create a link to a document in a different location.

Country Fact
Create a new Country Fact

Figure 11

Figure 12 Shows a graphical diagram for the document library described in Figure 10. This document library includes four content types: *Folder*, *Document*,

Link to a Document and *Country Fact* along with their columns. *Country Fact* is a custom content type whose content type hierarchy is shown in Figure 13. Note that in accordance with SharePoint inheritance rules any item using the Country Fact content type will include attributes Title (inherited from Item), Name, Managed Keywords (inherited from Document), Country, Information Category and Information Source.

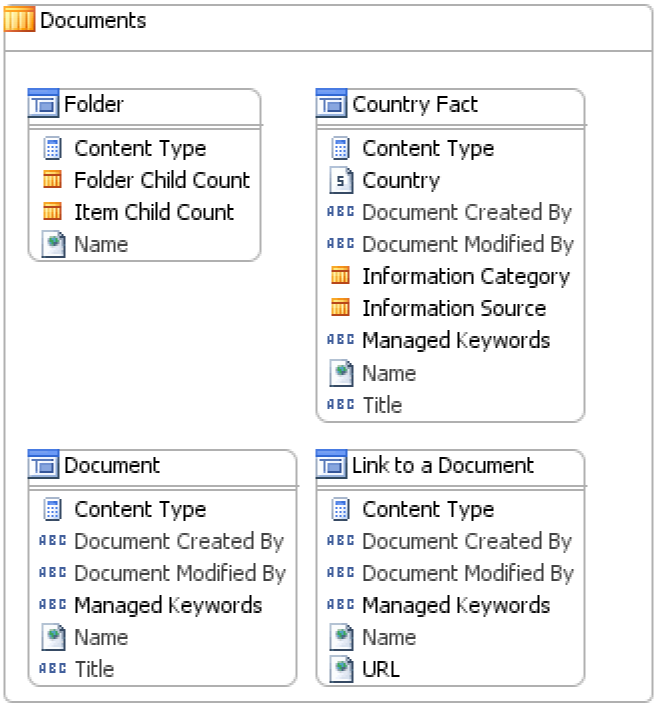


Figure 12

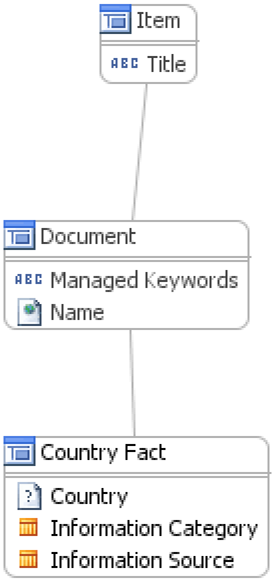


Figure 13

Content types defined on a site level provide an additional opportunity to structure content in lists in a way that will facilitate common search and content control practices. For example, site content types can be reorganized to reflect the structure of a company and its products and services.

In order to do this, similar columns applicable to many content types should be first extracted and moved up to parent content types. These can then be re-organized to reduce duplication, thus forming a corporate hierarchy of the content types. Next, the SharePoint site collection should be redesigned in a way that reflects the hierarchy of the company workgroups with clearly defined policies of how sites on each site hierarchy level can be created and managed.

Finally, content types should be positioned in the appropriate place in a site hierarchy for their intended use. This will promote creation of new sites that will inherit standard corporate metadata models, search and navigation techniques and general workflows.

Figure 14 Shows a graphical diagram of a re-based content type hierarchy where Country Fact content type is split in to two separate content types, Country Fact and Project.   Country Information is now a parent type for Country Fact can be used independently in any list within the current or child site. The use of this content type hierarchy would make it possible to assign a single information policy or to enforce certain rules for all documents using the Country Information, Country Fact and Project content types.

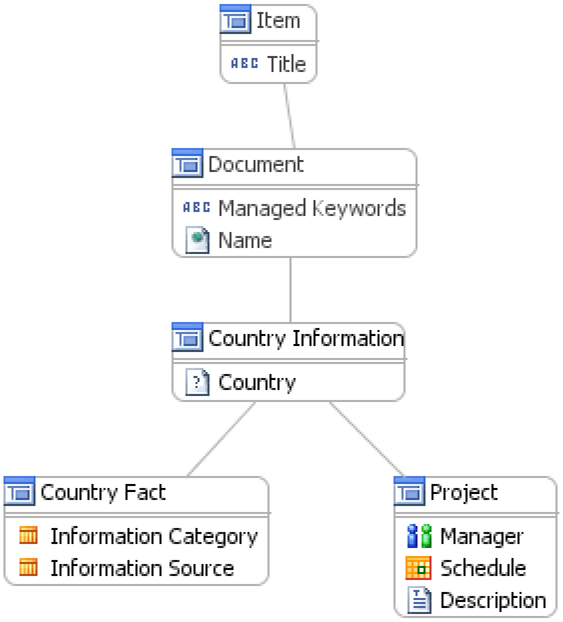


Figure 14

Occasionally the budget of an initial SharePoint implementation is large enough to analyze its intended use and thus design an appropriate site and content type hierarchies prior to roll-out. More likely, the design of a SharePoint site collection will grow organically along with the needs of an organization. The process usually evolves from the use of out-of-the-box SharePoint structures, to the addition of list level columns, to a reusable site level content type structure and a well-defined site hierarchies. The primary driver for each step in the progression is the lack of an effective search, navigation or workflow topology.

Part III

● *Classifying SharePoint Content to Improve Search and Control Content*

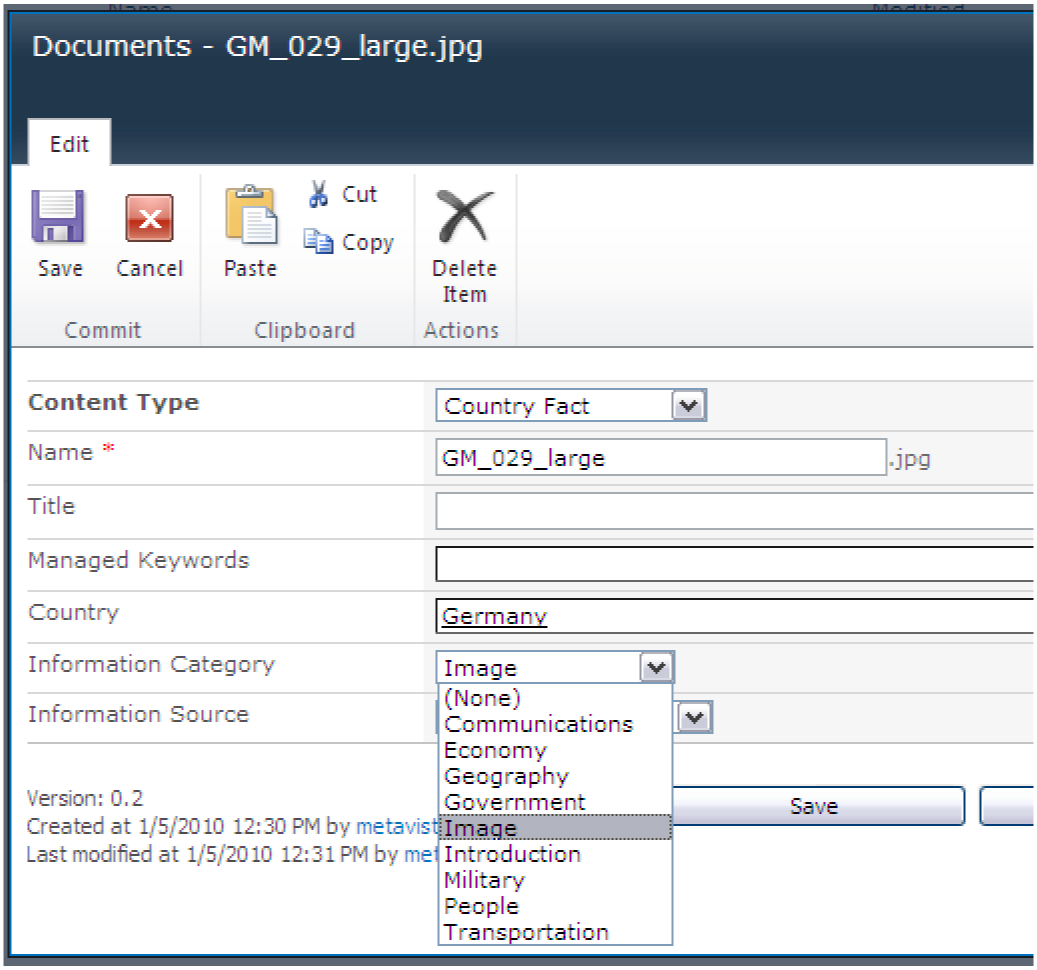
Document and item find-ability (and most workflows) in SharePoint rely on the actual metadata values associated with the content. Efficient metadata models improve search and navigation processes and enable generic workflows to be applied to content. This, however, will not be realized unless the metadata is reliably populated for all content. A metadata model architect (information architect) needs to achieve a fine balance between not enough metadata to make the content "find-able"; and excessive metadata which adds too much burden to users working with the content.

To encourage accurate and complete entry of metadata, architects should simplify the capture of the data for the user. For example, field values could be populated from a pre-defined vocabulary. Both the data entry control and the vocabulary could be configured to simplify the task. For instance, you could display a hierarchical tree of terms for the user to select from. Field values selection could be grouped in cascading relationships. Hence, the selection of a value in one field will limit the list of available values in another.

Authors are often unenthusiastic about populating metadata (despite having potentially written a large, time-consuming document). Nevertheless, the author is precisely the right person to do this task, since he/she usually has the most knowledge about the document.

The methods listed above do not address the issue of entering metadata values for multiple items already in the SharePoint environment that do not have any associated metadata; or files on a file store that are to be imported into a library (e.g faxes, scans, etc.).

Figure 15 Shows simple lookup drop-down box that helps users select a value for the field *Information Category*



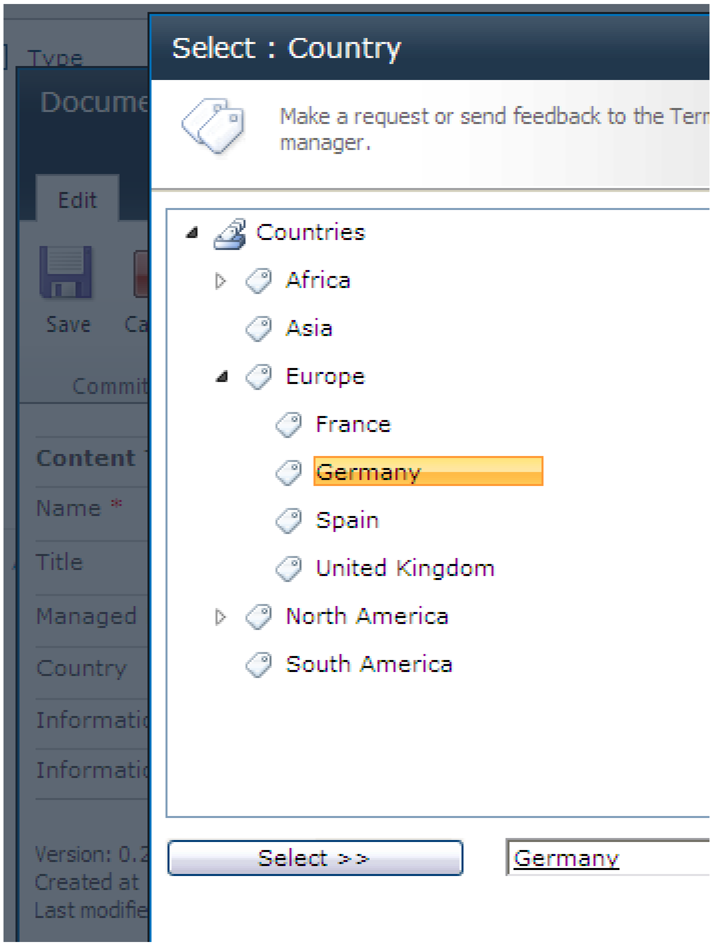
Documents - GM_029_large.jpg

Content Type	Country Fact
Name *	GM_029_large .jpg
Title	
Managed Keywords	
Country	Germany
Information Category	Image
Information Source	

Version: 0.2
Created at 1/5/2010 12:30 PM by metavist
Last modified at 1/5/2010 12:31 PM by me

Figure 15

Figure 16 Shows drop-down box with Countries taxonomy that helps user to select right value for the field Country. Taxonomy is a new feature of SharePoint 2010 that allows the selection of a value from complex hierarchical lists of terms.



Make a request or send feedback to the Terr manager.

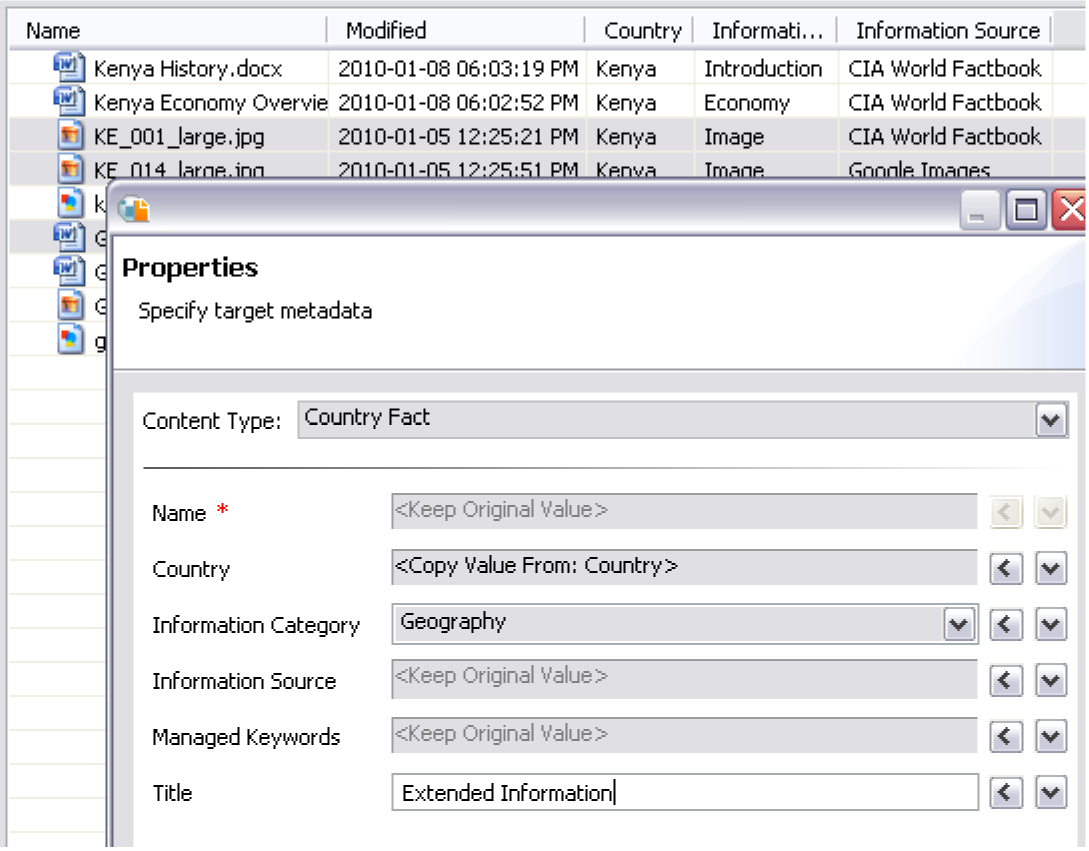
Figure 16

Several approaches are available for assigning metadata to a large number of items that have little or no existing metadata. One of these is to analyze the text of the document itself and then use an algorithm to extract and assign usable metadata values. These algorithms may be simple and rely on certain conventions that authors can adopt or they may be based on complex logic of specialized text analytics machine.

An example of a simple algorithm is to retrieve the 'Title' field from a cell in a spreadsheet. Text analytics strategies may be very complex and based on the dictionaries that require separate maintenance and time to learn data patterns and relationships. In some cases they produce reliable results with little human intervention. Considering the wide variety of content that can appear in documents and that many types of content (e.g. pictures, drawings, engineering models) have no text at all, successes in text analytics are rare.

Mass documents tagging could be seen as a reasonable balance between direct metadata entry and text analytics. Mass tagging involves filtering a specific set of documents based on a criterion that includes existing metadata and then updating the metadata values for all these items, en masse. Mass tagging can be done by typing a new value for a field, selecting value from a lookup list or mapping one field to another so that the current value from the source field is copied into the target. Site, list and folder locations of a SharePoint item are also part of its metadata set that could be changed during a mass tagging exercise. It provides a powerful way to relocate content based on its metadata.

Figure 17 Shows a third party mass classification tool that can change the content type and metadata values for many selected items and documents at the same time. Such tools allow organizations to gradually evolve SharePoint information architecture keeping pace with increasing user adoption.



Name	Modified	Country	Informati...	Information Source
Kenya History.docx	2010-01-08 06:03:19 PM	Kenya	Introduction	CIA World Factbook
Kenya Economy Overvie	2010-01-08 06:02:52 PM	Kenya	Economy	CIA World Factbook
KE_001_large.jpg	2010-01-05 12:25:21 PM	Kenya	Image	CIA World Factbook
KE_014_large.jpg	2010-01-05 12:25:51 PM	Kenya	Image	Google Images
k				
G				
G				
G				
g				

Properties

Specify target metadata

Content Type: Country Fact

Name * <Keep Original Value>

Country <Copy Value From: Country>

Information Category Geography

Information Source <Keep Original Value>

Managed Keywords <Keep Original Value>

Title Extended Information

Figure 17

Mass tagging, which can be referred to as metadata enhancement, relies on the pre-existence of some metadata. This can be as simple as a name, author and creation date that is auto created when the item is modified; or can be the result of a manual metadata assignment by authors; or an automated process such as text analytics. Mass tagging can be applied many times to different document sets, so that different aspects of metadata can be applied to documents selected by different filters.

- *Conclusion*

The need for a standardized content search and workflow dictate that metadata structures should be standardized and consistent across SharePoint environment in an organization. Different evolutionary approaches can be taken to design and maintain these structures. Coupled with multiple mechanisms for entering and updating metadata values for your SharePoint content will result in an effective, consistent and reliable search experience and an efficient automation of business processes through workflows.

About the Author:

Mark Klinchin directs the technological vision and product development for MetaVis Technologies (www.metavistech.com). Mark joined the company with 15 years of experience as a software product architect. As CTO of MetaVis he has led the development of MetaVis Architect Suite (www.metavistech.com/SharePoint) which takes the complexity out of designing, deploying and managing content within Microsoft SharePoint 2003, 2007 and 2010 by offering reusable taxonomies, metadata management and migration software and services. Mark holds Masters Degree in Applied Mathematics. You can follow him on twitter @mklinchin. You can download a trial of MetaVis Architect to see your metadata model at www.metavistech.com.

About MetaVis Technologies:

MetaVis (www.metavistech.com) provides software solutions to help organize SharePoint environments for improved search, findability and e-discovery. MetaVis takes the complexity out of designing, deploying and managing content within SharePoint by offering reusable taxonomies, metadata management and migration software and services. The benefit is an organized SharePoint environment that is easily understood and well documented.

The company believes that taxonomy management within SharePoint should not be complicated to implement and use. MetaVis products are based on intuitive, *graphical interfaces that are easy to use and easy to install. Drag and drop features allow information architects to design SharePoint metadata models and reuse them saving valuable time and resources. As a result, MetaVis products improve search optimization, consistency, content migration, and workflows across corporate SharePoint sites.*

You can follow MetaVis on Twitter @metavistech