

[\(\)](#)[about us \(/about-us\)](#)[services \(/services\)](#)[showcase \(/showcase\)](#)[perspectives \(/perspectives\)](#)[events \(/events\)](#)[contact \(/contact\)](#)

Thoughts and ideas
on things we
care about!

Subscribe to PebbleRoad
Press

Handpicked stories from the
team at PebbleRoad.
Delivered every week!

Metadata Collection Strategies

By [Maish Nichani \(http://www.pebbleroad.com/about-us\)](http://www.pebbleroad.com/about-us) | 14 May 2013

Categorized under:

[Tweet](#)

| [Comments \(0\)](#)

This article was co-authored with [Patrick Lambe of Straits Knowledge \(http://www.straitsknowledge.com\)](http://www.straitsknowledge.com).

Metadata is defined as "data about data". In data systems, the metadata describes the data resources and how they are structured. In other information systems such as content and document management systems, metadata describes the structure, content, and management rules associated with the content or documents. Typical metadata for a document might be 'author', 'title', 'subject', 'date of publication' and 'security classification'.

Metadata is important in enterprise settings. If you have a common metadata framework across all your systems, using common controlled vocabularies, and if metadata is consistently and reliably applied to your information and data assets, it can:

- connect information and data from different different systems
- help find and retrieve information more efficiently and effectively
- help better manage the information lifecycle (including the delicate balance between accessibility and security)
- support coordination and collaboration from a common information base
- build a sense of shared purpose

The challenge is to get complete, accurate and consistent metadata applied to information and data resources.

Metadata can be collected in many ways—from the information environment, work activities and from people. The problem arises when metadata that could be effectively collected from the environment is delegated to be collected from people. People who are in the middle of work tasks do not see direct benefits from completing numerous metadata fields. When coerced into doing unnatural things, they usually revolt or find workarounds thereby undermining the entire initiative.

In this article we share strategies to collect metadata that lower the reliance on people in supplying metadata. We cannot completely remove people from the equation but we can prevent them from doing additional work, and focus the role of people on the value added metadata that machines and environment cannot automatically supply.

Principles

In the course of our work with large enterprises we have found the following principles can be used to set the baseline for the conversation on collection strategies.

Collect (and expose) only what is required

Not all metadata is required all of the time. For example, in an organisation that deals with different industries, metadata for industry coverage may only be relevant to the externally facing departments, not support functions like Human Resources. Similarly, if documents are only meant for personal use or for use within a small team, they will require only basic metadata. Generally speaking, the need for additional metadata grows as the need to expose the

document to different audiences grows.

Collect incrementally

You don't have to collect all metadata at one time. For example, documents start out private to the individual and when ready it is shared with the team or the entire organization. The private document requires little metadata and all of it can be collected automatically. But when the document becomes sharable it needs additional metadata allowing it to be found and used. It is best to collect metadata required for a sharable document when the status of the document is changed from 'private' to 'sharable'. It would be annoying if all of the metadata is collected when the document is still in the 'private' state. Metadata is only added as needed.

Strategies

Here are some metadata collection strategies that can help in collecting complete, accurate and consistent metadata.

System-assigned values

A lot can be inferred in an enterprise setting. For example, an enterprise system will likely know the following:

- Name and ID of user
- Department, e.g. Sales
- Geographic location, e.g. Singapore
- Job function, e.g. Business Development

Such information can be automatically supplied by the information system in place.

At times, however, it is not possible to predict the actual values such as the actual author of the document. The person authoring the document may be the secretary or someone helping out the actual author. In such cases the author field should default to the currently logged in user, but the field should be editable. This makes it easy to change the author field in those few instances when such an action is required.

Document templates

Document templates are pre-structured documents used for common documentation tasks. For example, a 'minutes of meeting' document template helps staff to quickly write up the minutes of a meeting using a pre-structured template. Metadata values can be pre-assigned to document templates. Subject topics that help people search for the document (such as document type) are very easy to predict in templates. Some business activity subject topics may also be predictable from the template - e.g. a press release template could be assigned a 'corporate communications' topic. This way when staff use a template the relevant metadata is already applied to it, and the burden of assigning metadata values is reduced.

Similarly, within very clearly defined process flows, the 'parent' templates for that activity can be set up at the start, and the metadata properties for the activity, and the workgroups responsible, can be inherited for all the subsequent 'child' templates.

Document locations

Many information and document management systems offer a drag-and-drop functionality enabling staff to drop a document to a specific location (folder or library). The specific folder or library can be pre-assigned with relevant metadata (e.g. subject topics) and then applied to all documents filed there.

Business rules

The selection of some metadata values may imply the use of other values. These need to be specified as business rules. For example, if a metadata value, 'document type', is specified as 'budget', then another metadata option, 'business activity', can be assigned to 'financial management' using a business rule.

Tag bundles

Some people have very specific job functions. In such cases the metadata they supply is often in the same combinations. Tag bundles (<http://www.pebbleroad.com/perspectives/using-tag-bundles-in-intranets>) are an easy way to pre-assign relevant metadata values as commonly occurring bundles. For example, a staff member may work consistently with specific organisations and document types. Instead of applying individual metadata values again and again she can set up a tag bundle with the common organisation names and the document types (and maybe topics) that she uses. In future, assigning the bundle applies all the values in one go!

Automatic collection

Auto-classifiers have been around for some time now. They are still not accurate. But they are getting better! Solutions of this type work by analyzing information and then using natural language processing algorithms to extract entities that can then be mapped to metadata values. These systems work better with metadata values that describe specific entities (such as locations, organisations, people, things). They do not work so well with more abstract entities such as concepts or activities, where the language describing them can be much more variable, and not so easy to identify automatically. The only advice here is to do a proof of concept before purchasing such systems.

Conclusion

Having complete, accurate and consistent metadata is essential to any information system. Collection of metadata therefore becomes crucial and having the right set of strategies in place can make all the difference between an effective system and one that is a glorified file store. People can be ingenious at avoiding onerous metadata assignment tasks, especially if they are repetitive and not apparently related to the task at hand. People also see different aspects to a document that to them are relevant but may not be to others, and this leads to problems of consistency. Smart design of systems and the way metadata is collected, using the principles and strategies above, can mitigate the burden on people, and improve consistency and quality of metadata.



Add a comment

[Subscribe to PebbleRoad Press](#)

Handpicked stories from the team at PebbleRoad. Delivered every week!

PebbleRoad Pte Ltd, 314 Tanglin Road, Phoenix Park, #01-14, Singapore 247977

© 2013 PebbleRoad Pte Ltd

+65 6736 2180, info@pebbleroad.com (<mailto:info@pebbleroad.com>)