Enterprise Content Categorization – How to Successfully Choose, Develop and Implement a Semantic Strategy

KAPS Group
Knowledge Architecture Professional Services

Table of Contents

About the Author

Tom Reamy is the Chief Knowledge Architect and founder of KAPS Group, a group of knowledge architecture, taxonomy and text analytics consultants.

Tom has 20 years of experience in information architecture, enterprise search, intranet management and consulting, education software and text analytics consulting. His academic background includes a Master's in the History of Ideas, research in artificial intelligence and cognitive science, and a strong focus in philosophy, particularly epistemology. He has published articles in various journals and is a frequent speaker at knowledge management conferences.

When not writing or developing knowledge management projects, Tom can usually be found at the bottom of the ocean in Carmel, CA, taking photos of strange creatures.

Introduction

Effective information access has been a struggle for organizations for decades. Companies have tried giant enterprise search and content management initiatives and the situation just got worse. They tried adding metadata (badly) and things still got worse. They tried project after project, but things continued to get worse.

However, this situation has changed recently with the introduction of semantic technologies that add a new level of intelligence and meaning to enterprise content management. The concepts behind these new technologies are the key ingredient to success.

While these new technologies alone offer great promise in finally really improving an enterprise's access to information, these technologies also call for new approaches to get the full value from them. In other words, there is a right way and a wrong way to adopt these technologies.

The wrong way is to view this endeavor as simply another IT project that involves doing a standard requirements gathering; doing a standard software evaluation based on features and technical specs; buying and installing the software; implementing an application for whatever group is leading the project; and then declaring the project done.

What is wrong with this process is that it does not include an understanding of the role of language, or semantics, within the organization – and why that requires a different approach. It also lacks the strategic vision required to derive ongoing value – that is, to continually deliver benefit by ensuring text data is utilized as an informational asset that is part of the enterprise's semantic infrastructure. While it is possible to get limited and measurable value out of an isolated project, the way to get maximum value is to develop a strategic vision of how automating and associating text data fits within – and ultimately transforms – your organization. This doesn't mean that your first text analytics project has to be a giant, multi-year effort. But having this strategic vision does ensure that your first project – and your last – will continue to deliver maximum value, avoid dead ends and duplicate effort, and create the means to solve many information access problems.

A good way to think about a semantic infrastructure approach is to compare it with your IT infrastructure. An IT infrastructure creates a platform for applications to run on and enables those applications to communicate with each other. In the same way, a semantic infrastructure is a platform for building a broad variety of applications and enabling those applications to

communicate with each other – at a higher level than simply exchanging data. It has more to do with the network than specific uses of the network.

There are three dimensions to a semantic infrastructure:

- Content and content structure, such as metadata standards, taxonomies and other structures.
- Technology, which can include applications such as categorization, search and content management.
- A team of people dedicated to maintaining, refining and facilitating the application of the infrastructure's various elements.

None of this is particularly new or radical. The point is that your organization has a semantic infrastructure whether you consciously know it or not. The difference among organizations relates to the extent to which this text data is structured and how well understood it is.

A tragedy of modern organizations in this age of information and knowledge workers is how little effort typically goes into the semantic infrastructure. For example, most organizations have less than one full-time person for search functions, and librarians are seen as fluff. Employees at these organizations then complain about how they are swamped with information and have to spend so much wasted time looking for things.

When evaluating your current infrastructure and what you want your semantic vision to be, a few key questions need to be answered. Some examples of these questions follow.

- What is the full range of information or types of content within your organization, and who creates that content?
- What are the information needs for which the content is intended to be used, and how is the content actually used within your organization's business processes?
- What information technologies do you have now, and how are they being used? How might they be used in an ideal world?
- What kinds of information structure resources do you have now – such as taxonomies, file plans, dictionaries, metadata standards and implementation guidelines?
- Who developed and who maintains these information structure resources?
- What new information sources could your employees use to do their jobs better?

Answer these questions for your current text information processes and you'll be able to detail the three dimensions of your current semantic infrastructure: the content and content structure, the technology, and the team. Working through this exercise will also help define the vision of where you expect to be – as well as the priorities for how to get there.

Regardless of the form it takes, and the methods you use to develop it, a deep and accurate understanding of your organization's semantic infrastructure is the foundation that will support all your information and knowledge management initiatives. What you use your semantic infrastructure for is a platform upon which you can build and support a number of applications. This can range enterprisewide applications, like search or document management, to specific projects like a web content categorization application or a customer support application.

"Self-knowledge is the highest form of knowledge."

– Plato or one of those Greek philosophers

Semantic Infrastructure and Information Strategy

Before you can reasonably begin any enterprise text initiative, you must start with a fully articulated understanding of your organization, the strategic and business context within which you will develop the initiative, and, as indicated above, a deep understanding of your semantic infrastructure.

Martin White and others have written about the foolishness of trying to do enterprise search without a fully developed strategy document.[1] To do enterprise content categorization properly, it is even more important to start with a deep strategic understanding (i.e., self-knowledge). Whether you use a formal research effort to develop this strategy or approach it less formally, it is essential that you have this understanding. If not, you risk making a poor decision because of an inadequate understanding of what the technology can do. If you lack this understanding of the technology, you will likely find plenty of vendors out there waiting to take advantage of your unfamiliarity. The net result is usually the wrong technology along with inadequate development resources, and another wasted opportunity to actually do something about your information access problems. Consequently, it will likely be all the harder to invest again in such efforts for the foreseeable future.

[1] White, Martin. "EcontentMag.com: The Sorry State of Search Satisfaction." *EcontentMag.com: Digital Content Strategies and Resources*. Web. 19 Nov. 2010. http://www.econtentmag.com/Articles/Column/Eureka/The-Sorry-State-of-Search-Satisfaction-61565.htm

Creating this strategy includes developing a model of how enterprise content categorization can be used in your organization. To get there, some level of research is typically required to answer questions like:

- What are the information problems throughout your organization, and how severe are they?
- What are the differences between information problems in your various departments?
- Do you have multiple search engines? If so, what general capabilities and text analytics capabilities do they have?
- What is your current content management environment? Do you have multiple technologies and repositories, and can they incorporate enterprise content categorization functionality?
- What is your current publishing process, and how is metadata added now? How would you like it to be added?
- What kinds of search-based applications do you have now, and what do you envision for the future?

And so on.

The point is that you need this strategic and organizational foundation to make the right decision about how to approach an enterprise-level content categorization initiative; how to choose which software package(s) is right for your organization; and, even more importantly, how to do the actual development of the text analytics initiative.

Acquisition Strategy

Of course, before you begin the actual process of evaluating software, you also need to assemble your team. There are no established standards for this, but there are some best practices. First, the team for evaluation and development should be an interdisciplinary group with members from IT, business and library science, or some other sort of information professionals. Just as the process of information strategy development creates the foundation for future efforts, the evaluation team will also live on beyond the evaluation phase to become the development team.

Text analytics, especially enterprise content categorization software, differs from most other information technologies in that its core capabilities of auto-categorization, auto-summarization and entity extraction have more

to do with meaning and semantics than with technology. Furthermore, this software is not designed to be used by itself; rather, it is meant to be used with other technologies such as search and content management.

As a result of these key differences, the process of evaluating and selecting the best software for your organization should focus on two key areas:

- Development and ongoing maintenance of a semantic layer.
- Integration with other information technologies.

Types of Software

It was only very recently that this field of taxonomy creation and content categorization software became standardized. Now, most major vendors have settled on a core set of capabilities that includes the following:

- Taxonomy management.
- Auto-categorization and a categorization workbench that supports multiple, advanced techniques as well as the ability to test categorization rules on a variety of content.
- Entity extraction and a workbench that enables development of large catalogs of entities as well as categorization-type rules to extract novel entities that are not in the catalog.
- Auto-summarization.

In addition, some vendors now offer two additional capabilities:

- Fact extraction and a workbench to develop entity extraction-like catalogs and rules for dynamically extracting facts.
- Sentiment analysis and a workbench to identify and fine-tune the opinions expressed in content.

The vendor space is complicated by the fact that a number of search engine and content management vendors have begun to offer categorization capabilities as part of their development environment. There is also the rare case where all of these capabilities are offered in a single, comprehensive suite that not only includes the capabilities listed above but complements them with traditional text and data analysis reporting and integration tools – that is, text and data mining tools.

This leads to a matrix of decision points about what type of purchase to make.

- A standalone taxonomy management package.
- An enterprise search platform that may include some categorization or taxonomy management functionality.
- An enterprise content management package that may include some categorization or taxonomy management functionality.
- An enterprise content categorization package that includes all the categorization and taxonomy management capabilities.

Standalone Taxonomy Management Software

Taxonomy management software is used to manage the creation, development and deployment of taxonomies. Its core capabilities are basic editing features, support for taxonomy-standard formats like SKOS, import-export formats, security and role-based access rights, support for multiple contributors, and other basic software features.

Standalone taxonomy management software is primarily used in situations where there are multiple taxonomies or very large taxonomies with many distributed contributors. More and more taxonomy management vendors are offering additional capabilities such as categorization. So the initial question to consider is: if you think you'll need text analytics, then determine if the text analytics software's built-in taxonomy management capabilities are sufficient for both your current and longer-term taxonomy management needs.

Enterprise Search Software

Some search engine vendors offer a subset of categorization capabilities. Typically, these include some type of categorization and entity extraction. Having these capabilities in your search product is good – especially as a hook for integrating more complete content categorization capabilities – but search software is not well-suited for developing those capabilities. Typically, enterprise search technologies categorize through the use of training sets, but they are rarely sufficient for good categorization that goes beyond general, high-level types of categories like "news," or "sports."

In other words, the categorization capabilities found within search engines are helpful, but are not enough for advanced categorization and extraction

functions. This limits their usefulness, and so they have to be supplemented by one of the other options.

Content Management Software

Content management software vendors added general categorization capabilities a number of years ago, but most of them were used to help create and manage keywords and other metadata like author, publisher and so on. Having this sort of metadata tool as part of a content management system is definitely a plus, but like the built-in capabilities in search engines, these options are not as advanced as those typically found in an enterprise content categorization platform. If your categorization and metadata needs are quite limited, this might be an option. But if you need (or can benefit from) more advanced categorization and extraction capabilities, then you will want to supplement any built-in functionality of a content management system with a dedicated enterprise content categorization platform.

Still, like the situation with search engines, having some built-in categorization and metadata functionality can make it easier and more effective to integrate content management (and search) with enterprise content categorization.

Enterprise Content Categorization Software

Advanced text analytics calls for a dedicated, stand-alone enterprise content categorization platform.

To develop the most advanced and useful text analytics capabilities, the best choice is a dedicated, standalone enterprise content categorization platform. This type of platform gives you a full set of capabilities, including categorization, entity and concept extraction, and summarization. In addition, you can add de-duplication, fact extraction and sentiment analysis as needed.

This platform gives you the most complete solution for developing content categorization capabilities – well beyond those typically found in taxonomy management, search and content management software. In addition, the built-in taxonomy management functionality of enterprise content categorization software is usually more than sufficient for the smaller taxonomies that are typical of categorization applications. The reason is that taxonomies used for categorization are typically not highly complex and do not have too many levels. Instead, they have categorization rules associated with each node, which avoids the need to rely on multiple levels

of a formal taxonomy. Finally, the technical issues around integration of content categorization with content management and search are relatively simple and well-understood, so there is no reason to rely solely on the limited categorization capabilities built in to search and management applications.

The Software Evaluation Process

Evaluating this type of software is best done as a two-part process. The initial phase is to take a relatively quick look at a wide number of vendors, keeping in mind your information strategy and semantic infrastructure vision. This review consists of an evaluation of each company, its reputation and market strength, and its high-level feature set. Maintaining a clear, longer-term vision – even when a specific pilot project is identified – will force this initial evaluation to include factors like extensibility, flexibility, customization and investment. This first look also acts as a filter to limit the number of companies that you will invite to hear your vision and immediate priorities. In turn, you will only need to ask a few vendors to present their software so that you can get a sense of how it operates.

The next phase of the evaluation is the heart of the process, and it entails doing a short, four-to-six-week proof of concept (POC). During this POC, each vendor will install their software and work with a team of experts – taxonomists who have experience with categorization and entity extraction, and subject matter experts (SMEs) who have knowledge of the domain of investigation – to test the software in a real-world environment with actual content.

Conducting a Proof of Concept

The only way to actually compare the quality of categorization and entity extraction is by developing real-life scenarios, real-life categorization and entity extraction, and real-life content. The reason such a process is necessary is that the fundamental difference between software packages is the quality of the results of categorization and entity extraction – which has very little to do with traditional software differentiators and everything to do with semantics and the meaning of individual words and their context within larger documents. The reason it needs to be a full POC is that this type of evaluation will help you determine how well the quality of the results can be improved by refining the categorization rules.

Even though this evaluation process can be somewhat lengthy and resource intensive, it is the only way to get an accurate picture of how the

different categorization and entity extraction technologies will perform in your environment. It has the added value of giving you a head start on your taxonomy and categorization development.

Benefits of a POC Evaluation

As explained, the first benefit of a full POC evaluation is that it can reveal how the software really functions in practice. Real-life tests almost always produce better results and also tend to uncover unexpected problems and even unexpected advantages. For example, in one project,the documents were so long that we had to develop different types of rules that combined the beginning and end of the documents, while discounting all the vast text in the middle. If we hadn't done this, every document would have been classified in every category – hardly a useful exercise.

A second benefit is that this POC process will provide good feedback on your current and projected taxonomies. This feedback can help determine how suitable the taxonomies are for categorization-based applications. Problems can include structural issues, like mixing general and specific terms at the same level of the taxonomy – or it could be that taxonomies that are fine for indexing may not work very well for categorization.

A third benefit of the POC process is that it provides a good idea of how the advanced Boolean categorization rules work, which is something you can't accomplish with demonstrations or documentation.

A fourth benefit is that the POC provides good information about the process and ease of customization. All software usually has to be customized during implementation – and this is particularly true for semantic-based software. If you understand how customization works and how difficult or easy it is to do, that will be a major differentiator between software packages.

A fifth benefit is that you will likely have metrics that indicate the type of efficiency improvements that will be gained by continuing this work. For example, you will learn the extent to which manual processes are reduced, the impact of more relevant content being retrieved and so on.

Outcome of the POC Evaluation Process

The final outcome of the POC process will provide enough information for you to make a business justification decision about which software package(s) to purchase.

The final reason for conducting a full POC process is that your implementation process will have already begun. Consequently, you should be able to make significant progress in terms of understanding the software; creating the best taxonomies and categorization-entity catalogs; developing a faceted navigation application utilizing categorization and entity extraction; and establishing which components are needed for solving your organization's information access issues. Last, you will have a team that has already been introduced to the necessary development activities.

Development Processes

The process of developing a semantic infrastructure should build on the POC, using the knowledge and experience gained from that exercise as well as the same team, approach and resources. If you were not able to conduct a POC with your data and resources prior to this stage, it is recommended that you begin with a pilot project for this aspect of the development process.

Applying categorization rules to a document produces a relevancy score.

The overall process of development is essentially a continuation and expansion of the POC process, with added attention to the set of applications that will be built on the platform. These applications will influence such things as the level of precision needed and the overall facet design(s) that entity extraction will support. While the applications will vary from organization to organization, the actual development process has common characteristics. Each step of the process is described in more detail in the following sections.

Categorization

The first and usually most difficult step in development is categorization. When thinking about categorization, it is important to remember that we are talking about partial membership in a category. To put it another way, applying categorization rules to a document produces a relevancy score just like a search engine. It happens to be a much better relevancy score than search engines because of the superiority of categorization over simple search – but still, when a categorization rule is applied to a document, what you typically get is a set of scores that indicate the likelihood that the document belongs to a certain category. And when a whole set of categories or a complete taxonomy is applied to a set of documents, what you get is a series of scores for each category, with a relative score for each document. Whether or not a document "belongs" to a category is determined by selecting cutoff values – and this can be quite flexible.

Preliminary Steps

Before you can begin developing categorization, you need two things. First, you need a suitable taxonomy. This taxonomy doesn't need to be huge with multiple levels. This is because categorization taxonomies are typically designed to be used in conjunction with other information elements like facets, and they are rarely more than three levels deep. Second, you need good sample content that can be used to develop categorization rules. Ideally, at least some of this sample content should be categorized by humans for both training and testing purposes.

Training documents are typically used for developing the initial set of rules. Testing documents are used to expand and refine those rules by validating against more and more content. This content does not necessarily need to be categorized by human subject matter experts, but it is helpful to do it this way.

Iterative Categorization Development

Step one: first round of categorization rules

Once you have the content, the next step is to develop basic terms that appear in the selected training documents. This can be done by hand with the taxonomist simply reading five to ten documents and looking for important words. It can also be done by the software, using an auto-suggest feature,[2] or it can be done with a variety of linguistic tools to extract words and phrases from the document. The first round of categorization rules are typically a set of words that are quite rudimentary. This first round also begins the develop-test-refine cycle (illustrated on page 13).

Step two: initial rounds of develop-test-refine

This initial round is best done by the development team only, with no real participation by any SMEs. The initial round consists of testing the first set of rules against new content and analyzing the results to see where the rules are weak. Rules can be weak because they miss content that is known to be a good candidate for a certain category, or because they miscategorize documents.

[2] Terms for training documents can also be identified by some text mining applications, wherein the software identifies different topics that emerge from the text collections.

The next step is to refine the rules. In some cases this can be done by adding or deleting terms, but it's important to use Boolean operators to create more sophisticated rules. These rules might be simple, such as "only count terms in the first 100 or 200 words of a document" or "only count a rule as a successful hit if it contains at least three occurrences of the word." They may also be more complex, containing operators such as DIST or SENTENCE, which indicates to only count terms that appear close to each other. Another typical use of these rules is to eliminate terms that appear in all documents because they are formal elements, such as a set of navigation labels.

A common pattern in this and subsequent rounds is to add more and more rules and terms that appear in the training set until they capture all the training documents. Then, you can start pruning the list of rules and terms so they become more general and can be applied to new content.

One question that often comes up is whether it is better to develop a single category to some level of precision and then go to the next; or whether you should develop a bigger set of categories to a low level and then go back through and refine each category. The answer is – whatever you prefer.

Step three: multiple rounds of develop-test-refine

The next step is to repeat the develop-test-refine cycle until testing gets to a level at which any further refinement needs SMEs to evaluate the results. One typical activity in this third step is to expand the testing corpus and test against a wider and wider set of documents.

Step four: multiple rounds of develop-test with SMEs – and refine

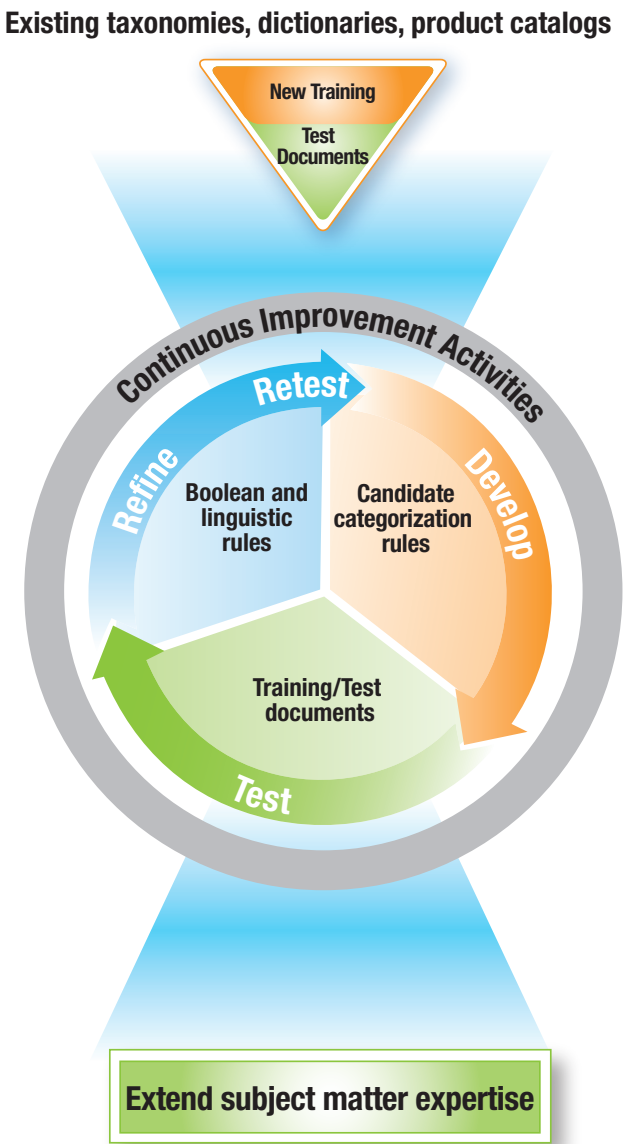
Once the categorization taxonomy has reached a level of precision on a wider set of documents than you had in the initial training corpus, it is time to bring in human expert feedback, the SMEs. This step can be implemented as a formal scoring procedure, or a more analytical evaluation of the quality of the rules.

Step five: repeat until "done"

As you might imagine, what constitutes "done" will vary from organization to organization and application to application. You may want to achieve 95% accuracy before you start using it in production, or you may want to go ahead with 80% accuracy and employ feedback from your users to improve

accuracy. There are costs and benefits associated with both approaches, and there is no correct answer. It depends mostly on the culture of your organization. If you think there will not be too negative a reaction to only 80% accuracy, then that approach might be a cheap and fast way to get a lot of feedback and also some early benefits. On the other hand, in some organizations aiming for only 80% accuracy might generate so much negative reaction that the success of the application will be permanently impaired.

The Development-Test-Refine Cycle



You can create viable, valuable taxonomies using the develop-test-refine cycle, but continuous improvements require multiple iterations of the cycle – including ongoing maintenance – to ensure that categorization remains in sync with new content.

Entity-Concept Extraction

There are two basic components, or technologies, for entity extraction. The first is a catalog of entity values, such as a list of people, organizations or products – Named Entities. The second is the development of dynamic rules to extract unnamed entities. Developing entity (also known as concept) extraction capabilities is significantly different than that for categorization, although there is some overlap in the development of dynamic extraction rules. Entity extraction development includes the following elements: designing the facet or class; finding and converting database content; building dynamic rules; and identifying development issues.

Designing the facet or class

The first step in developing extraction capabilities is determining what types or classes of noun phrase extraction (or concepts) you are going to need to design. There are certain standard classes (which will likely become facets in a faceted navigation application), such as people, organization, products and dates. However, for an enterprise content categorization initiative you will need many more – and more than you initially think.

For any one application, especially a customer-facing application like a product catalog, the specific set of facets will be largely determined by the properties of the products you store in your database. But it is a good idea to use application development as the impetus to research your users to determine which facets or characteristics of your products they find most important.

However, developing an enterprise extraction capability requires taking a broader look at all kinds of entities a wide variety of users or employees finds important or useful – and this should include both structured and unstructured content. Doing this requires significant research into user behavior and needs. Even for a product application, you can often add significant capabilities by including unstructured content that is outside the norm; in other words, content which requires extra research and design work.

Finding and converting database content

Once the initial facet design has been developed, the next step is to populate the various facets or classes of entities starting with whatever existing databases or catalogs you have in your organization. This might be an HR database of employees for a people facet, an organization chart

(hopefully dynamic) for organization names, a product catalog, existing dictionaries, and so on.

Turning these resources into extraction catalogs is usually the primary responsibility of a programmer rather than a taxonomist. The task is simply to write scripts or programs to convert the information in your existing databases into a format that your enterprise content categorization software can use as an extraction catalog. Of course, it is never simple and there are a number of issues that must be dealt with.

Building dynamic rules

In addition to converting databases into extraction catalogs, the other, more difficult task, is to develop rules to extract un-named entities. These rules are similar to the categorization rules; but instead of terms, the rules use grammar – such as parts of speech – and other variables, such as capitalization.

For example, a simple rule might be to look for a word starting with a capital letter and followed by a single capital letter; or, to look for a capital letter followed by a period and another word that begins with a capital letter. This would be designated as a likely candidate for a person's name. Another rule might be to look for a word beginning with a capital letter in front of the word "said" – because most of the time the entity saying something is a person. Another type of rule can be used to extract privacy information by looking for a person's name alongside a pattern of Social Security numbers: xxx-xx-xxxx.

Developing these dynamic rules follows the same process as that used for categorization – get sample content, develop initial rules and then follow the develop-test-refine process (illustrated on page 13) until the results are satisfactory.

Identifying development issues

Many development issues are associated with entity extraction.

The first is the same as the primary issue for categorization, and that is overall accuracy measured by precision and recall. Do the catalog and dynamic rules find all the instances within test documents? How many false positives do they find?

In addition to accuracy, another issue is scale. The number values for different entity classes can be in the millions. For example, one vendor offers an 800,000-named people database and 400,000-named organizations database. And there clearly could be much more than that. The scale of the possible values poses a number of challenges for developing and maintaining these databases or catalogs.

The other major issue is disambiguation. This is the issue of distinguishing named entities with the same name. For example, Ford can be a person, a car and a company. Distinguishing the different entities in text can be challenging, but with the addition of categorization-like rules, it is possible to do a good – though not always perfect – job. And finally, a related but simpler issue is when an entity has more than one name. This issue can usually be handled with simple synonyms,[3] or co-referencing.

Combining extraction and categorization

A final development element is how to combine extraction and categorization. A good software package should include this capability. For example, it should be possible to incorporate entity extraction with categorization rules so that the total number of entities, or entities for a particular class, can be taken into account by the categorization rule. In addition, as we have seen, extraction can incorporate categorization rules to help populate dynamically generated values from documents.

One important advantage of combining extraction and categorization is that it allows you to separate long lists of terms and variant spellings and keep them in dictionaries or concept files that will simplify maintenance of rules and lists. In addition, if the application is going to be extended to multiple customers or domains, the separation of customer or domain-specific vocabularies into separate structures limits the amount of customization that is needed. This approach means the rules will stay the same, so the only thing that will change is the text in a concept "dictionary."

Summarization

Summarization is developed very much like categorization, but is relatively easier to do and tends to be more dependent on the nature of the content. Summarization is usually aimed at individual documents, providing a paragraph-sized summary – although it can be longer for really large documents. It is important to remember that a summary is not what a human would produce – new sentences that try to capture the overall

[3] Note that development of synonym lists can be partially automated by including text mining capabilities as part of your enterprise content categorization implementation.

ideas of a document. That capability awaits the development of artificial intelligence, which is something many of us have been waiting on for more than 20 years.

What a summarization rule does is create a set of procedures that enable the software to select important sentences in the document and then put a number of them together to suggest the document's important ideas.

Analysis of the documents

The first step in developing a summarization rule is to do a thorough analysis of the document collection against which the rule will be applied. This is particularly important for summarization since the rules will depend on such variables as average length of documents, presence or absence of standard headings and subsections, and even overall quality of writing.

Initial rule development

Once you have fully analyzed the document set, the next step is to create a basic general rule to specify the important parts of the document and to specify variables such as length of the summary. For example, the opening paragraph in a well-written document is often where the major ideas are expressed. Similarly, the last paragraph is often just as important. This is where you can specify the use of document headings and settings, or even metadata fields if they are present.

Rule refinement

The next step is to apply the rule to test a training set of documents and to analyze the results and begin the iterative process of improving the rules. There are very few standards in this area, so what constitutes an acceptable level will vary with the intended application. This is something you should have discovered during the initial research or strategy phase that preceded actual development.

Summarization and categorization

One of the common uses of summarization is to provide a search results list that summarizes content so users won't need to open each document. The summarization rule needs to take a search query and focus the summary on terms of the query that can't be known ahead of time. The technique that an integrated enterprise content categorization software package employs is to build categorization rules into the summarization rules. This

Summarization provides abridged content so users won't need to open each document.

technique enables the rule to use a query term along with an associated categorization rule to select not just the generally important parts of the document, but also the parts where the expanded search terms appear.

Sentiment Analysis

Developing sentiment analysis is very much like developing categorization rules – but there are some significant differences, based on what you are looking at to build your rules. As with categorization, you typically start with a training set of documents that can be used to generate an overall sentiment score. But statistical scoring is only the first step.

Entities and components

Rather than setting an overall positive or negative tone for an entire document, sentiment is normally used to capture positive and negative sentiment about a set of entities and components – like a set of cameras and their features, such as lenses. So the first step, as with entity extraction, is to design what it is you're going to track – that is, what entity classes and what components or features you're going to drill into.

The process then follows the same process as categorization. First, you select a training sample of pre-classified positive and negative documents about each entity class or components you are interested in. You then use the software to suggest terms that are positive or negative, and build sentiment rules from that. The next step is to refine those rules using the same develop-test-refine cycle that is used for categorization.

Types of terms for sentiment rules

Although the process of developing sentiment analysis is the same as for categorization, the types of terms that go into the rules are very different. Sentiment is typically expressed with phrases, verbs, adjectives and adverbs rather than nouns, which are often used in categorization rules.

Development issues

In some ways sentiment analysis is more difficult than categorization – but in other ways it is easier. It is more difficult because you have to determine what a document is about before you can determine what the sentiment

is. So sentiment builds on top of categorization. But since sentiment is more often associated with entities than with subject categorization, it relates not so much to what the article is about, but to whether or not the sentiment being expressed is about a particular entity. And this tends to be a simpler task.

Perhaps the most difficult aspect of sentiment analysis is distinguishing tone and attaining a deeper understanding of the context than is required for categorization. For example, sarcasm is particularly hard for software to understand. So the phrase, "Yeah, that camera is really great" can be either positive or negative depending on the surrounding context. Sometimes certain code words assist with the interpretation – like "Yeah, ***right***, that camera is great." But often, those code words are not available.

On the other hand, sentiment can be simpler than categorization because it only requires dealing with what is, in essence, a three-node taxonomy coupled with some entity extraction. Determining whether a statement expresses a positive, negative or neutral sentiment can be much easier than trying to determine which of hundreds of categories are being discussed in a document. However, this simplicity is offset by the fact that you typically need to score the various positive and negative sentiments expressed in the document about a wide variety of topics or entities rather than only needing to score the document's overall sentiment. For example, in a common sentiment application that involves tracking sentiment about products in online forums, you would create a small taxonomy of products and features such as phones -> particular phones (iPhone) -> and then a number of features, such as price, quality, battery, screen, etc. This taxonomy allows you to track the specific aspects or features of a product that are generating positive and negative buzz.

Maintenance and Refinement

In keeping with our theme that semantics is infrastructure rather than a project, it is important to remember to plan for maintenance of your taxonomies and catalogs. As new content is generated or acquired within the enterprise, and resources are extended to new external content repositories, or new applications for them are developed, your taxonomies and catalogs will need to be refined, expanded and updated. And just as SMEs are not the best source for developing categorization taxonomies, they are not the right source for maintaining and refining them either.

A dedicated taxonomy team is the best answer, although the team can be as small as a single, part-time resource for very small companies. For larger companies, one or more full-time taxonomists should be dedicated to this task. Many companies seem to balk at assigning taxonomists to ongoing tasks, but as we discussed in a companion paper (*Enterprise Content Categorization – The Business Strategy for a Semantic Infrastructure*), that is a very shortsighted attitude that will cost much more in the long run.

In addition to refining taxonomies as new content is generated, this dedicated resource must also work with application teams to incorporate the taxonomy with new applications. In other words, a taxonomy team should not be thought of as an isolated group; instead, it should be part of every activity throughout the organization that relates to improving access to information. Information access is improved either by enhancing platform applications – like search and content management – or by developing new applications that utilize information to improve everything from job performance to creating new products for external or internal use. Again, this is part of the infrastructure, the workings of the organization.

To perform this enterprisewide function, the taxonomy team needs to be in communication with all groups in the enterprise, just as a corporate library. And like a well-designed library, this communication is not just one way, with librarians asked to help find this or that information. Rather, the taxonomy team should reach into every department and every group to proactively determine how they might improve processes and applications. Another function of this group is to facilitate and educate others about the various ways these taxonomies can be utilized.

Where this team is located in the organization is not as important as its interdisciplinary structure. However, that interdisciplinary structure means that it should be either an independent group (such as a text analytics center of excellence), or part of another interdisciplinary group such as a knowledge management (KM) team. Fitting a taxonomy group within a KM group can be tricky given their different perspectives on knowledge and information. But in reality, merging the two can produce great benefits for both. Taxonomists can benefit from being integrated with the existing corporate structure, from the enterprisewide mandate of an existing group, and by gaining a deeper understanding of knowledge. The KM group can benefit by enriching their sense of knowledge with rich semantics, by having a semantic platform for their standard initiatives – like communities of practice and expertise location – and by adding a technological element with enterprise content categorization that fits better than other IT technologies.

Metrics and Feedback

To perform these functions, it is imperative for robust feedback mechanisms to be incorporated with all enterprise content categorization applications. This feedback should be a combination of passive and automatic feedback built into the software, along with active, user-generated feedback gathered through surveys and other techniques. These other techniques include social media software and so-called "folksonomies" that can enrich existing text analytics taxonomies with candidate terms as they uncover whole new topic areas.

Enterprise content categorization software can go way beyond the feedback analytics of enterprise search software. Instead of simply tracking specific search terms, enterprise content categorization software can use the whole rich set of semantic relationships built into the text evaluation structures. For example, instead of simply being able to generate a list of the most or least frequently used search terms, this software can track complex subject areas at any level of specificity, and uncover what is missing. In other words, if your semantics are well constructed, they can tell you what subject areas are completely missing – something that a list of entered words cannot do.

These kinds of metrics can be used not only to refine and improve your content categorization structures, but to potentially open up a whole new set of tools for understanding your organization. The metrics can reveal things such as:

- What are the important subjects that employees are looking into?
- What information do they need for their jobs?
- What information are they having trouble finding?
- How are those subjects related? This may include ways that could otherwise be missed in our specialized, siloed world.

Enterprise Content Categorization – Best Practices

This is not the place for a complete tutorial about how to develop enterprise content categorization projects, but I want to present some basic best practices to help people think about the best way to approach this task. We've talked about how to do the actual development, so let's take a look now at how to approach the project itself.

Process or Project Plan

There are two key aspects of the overall approach to the project:

A realistic budget

In a world where library services are seen as a luxury and metadata is seen as an afterthought or something "nice to have," it is important to realize that enterprise content categorization requires a real budget if it is to be done right. I have seen clients who don't blink an eye at spending a million dollars on yet another enterprise search engine, but who consider US$200,000 for library or taxonomy services (that will actually make the search work) too expensive.

A flexible project plan

The other key idea is that when you are dealing with semantics and meaning, you will need a more flexible project plan than with most IT projects. Part of this is because of the inherent complexity of language, and part of it is because there are fewer clear-cut, objective measures of success for this type of project.

The complexity of language means it is very difficult to predict just how long it will take to solve specific issues. And the complexity of objective measures means that it is difficult to determine when the project is finished and heading into the maintenance phase.

One way to deal with this complexity is simply to decide that the development phase will go on for a specified length of time and that the project will then enter the maintenance phase. This actually works fairly well with two caveats. First, you still need to be flexible about the timing for achieving completion of some of the sub-tasks. And second, it could mean that the maintenance phase will require ongoing development to achieve satisfactory levels of accuracy. In other words, the transition from initial development to maintenance is a fluid boundary that in the real world is often marked by the transition from using outside consultants to using internal taxonomy resources.

Resource Teams

In the area of resource teams there are three major themes – the need to have an interdisciplinary team, the critical element of communication,

and how and when external expertise can be best leveraged. Each of these is outlined below.

Interdisciplinary teams

As mentioned previously, enterprise content categorization requires an interdisciplinary team with significant involvement from a library or information (not IT) professional. Ideally, this team will be in a dedicated department if the organization is large enough.

However, it is not as important to have head count as to make sure different functional areas are involved. These functions include IT support and development, as well as business SMEs. These types of personnel provide input into the design of applications, and serve as a source for understanding the information behaviors that are part of any business processes. Finally, the team should include an information specialist (taxonomist, library professional, data scientist, etc.) who designs and develops the information infrastructure.

Importance of communication

Communication deserves special attention because of the interdisciplinary nature of enterprise content categorization projects. Project communication is always important, but when you have people in such different worlds using such different "languages" as IT, business and taxonomy, then it is even more important. And it involves more than just having regular project updates or meetings; it involves a translation function to make sure everyone understands the others on the team. Fortunately, if you have a taxonomist on the team, they should have already have experience in dealing with multiple organizational languages.

Merging internal and external expertise

Because most organizations don't have internal taxonomy resources when they begin these initiatives, most projects will involve the use of external consultants. Since I am one of those external consultants, I might be biased, but for now it seems pretty clear that this is true. It is important to remember that taxonomy creation and content categorization require more than standard library experience. Librarians are great resources for any text-related project, but they need to be trained in the differences that enterprise content categorization involves.

Merging internal expertise with external consultants is not difficult if the external consultant has experience, but it is something that needs to be included in project plans and designs. Aside from communication issues, the main function these external consultants provide is training internal resources how to best leverage the technology so that they can take over the ongoing refinement and maintenance of taxonomies, categorization and so on.

Categorization/Taxonomy Structure

There are a number of best practices and important themes to keep in mind when developing enterprise content categorization applications.

Taxonomy: tradeoff of depth and complexity of rules

Generally, you can achieve similar results by either developing a deeper taxonomy to achieve greater specificity, or by developing more complex rules associated with a shallower taxonomy. Which is better will depend on the application. For example, you could have "teachers" as a subcategory under "education" with its own rule, or you could simply include rules that specified teacher rules within the set of rules for education.

Categorization accuracy will vary depending on the content.

Multiple avenues: facets, terms, rules, etc.

There are always different ways to achieve good results through combinations of facets, terms and simple or complex rules. Again, the best answer will depend on the application. For example, you might have "people" as a separate facet that is broken down into subcategories, or you might include "people" as a subcategory under a number of higher-level categories like "education." You might also have "people" as one facet that is organized alphabetically, and allow users to combine that facet with other facets like "discipline" – which, in turn, is broken down into fields like "education."

Recall-precision: application-specific

It is also important to remember that accuracy, as measured by recall and precision, will vary depending on the content. If you know your content won't change much, then it makes sense to go for the highest accuracy you can for that specific content set – and you do that by creating the most specific rules possible.

However, if you have content that changes significantly or often, it is sometimes better to develop more general rules. While that might not be as accurate for a given set of content, this approach will achieve higher accuracy against significantly different content.

Training sets – like "find similar" – need rules

As mentioned above, training sets are a good place to start a categorization initiative. But you will never get a very high accuracy (>75%) with training sets alone, unless your content and taxonomy are both very high level and have very large differences between the categories – like politics and sports. Trying to refine and maintain training set-based categorization, particularly for new content, is very labor intensive.

No best answer: needs custom development

As a number of themes covered in this paper have implied, there is no best answer (almost no matter the question) when it comes to the development of an enterprise content categorization capability. This means that your project will require custom development. Beware of those vendors who promise that all you need is a standard, out-of-the-box solution.

Not all taxonomies are equal, and all require customization.

With simple taxonomies that do not have categorization rules, you can get value out of standard or prebuilt taxonomies. But even then, you probably need some customization to reflect the language of your organization.

There really aren't any standard or prebuilt categorization taxonomies – although in some cases, you can find starter categorization taxonomies that may (or may not) help reduce the time spent in the POC phase.

Technology

Basic integration

Integration is actually the easy one. Most enterprise content categorization software produces an XML file that any other program, search or content management application can read and incorporate. The basic output is often a list of files that have been categorized or populated facets that support faceted navigation. This output can be directly incorporated into a traditional standalone search results list or as different facet results.

Advanced integration

Faceted navigation combined with taxonomies is extremely valuable, but there are a number of ways to get even more value from integrating enterprise content categorization with other enterprise applications.

In a companion paper, *Enterprise Content Categorization – The Business Strategy for a Semantic Infrastructure*, we discussed using content categorization to enhance content management by suggesting metadata values for fields such as subject or keywords, and for various facet values, like people and organizations. Search engines can also use categorization output not just to provide ways to refine search results, but also to directly influence relevance ranking algorithms. This can be accomplished in two ways. First, it can be done through the use of taxonomic expansion of search terms with synonyms and related terms. It can also be done by using the categorization rules themselves to incorporate rules that, for example, tell the algorithm to only count the term "pipeline" if it is near "oil" or "gas" but not near "research."

This is a new area that requires a joint effort between IT experts and taxonomists. When you add the capability of integrating enterprise content categorization to data that uses semantic technology, the possibilities are staggering – because you can now combine unstructured content and structured data to create new kinds of applications that leverage the advantages of each type of content. The development issues here can be complex because they involve both IT and semantic contributions, and each project will call for new solutions and approaches. It is too early to tell what will emerge as best practices in this exciting, new area.

General Risk Factors

Most of the general risks associated with enterprise content categorization – and in fact, from semantic infrastructure development in general – arise out of the communication issues between application users, business groups and IT. Communication is always important, but when you are communicating about communication or semantics, the possibilities for misunderstanding are even higher because there is often no common framework on which to base discussions. Some of the typical issues that arise are briefly discussed in the following sections.

Value understood, but not the complexity and scope

Very often business and IT groups understand that enterprise content categorization has the potential to greatly aid information access. As a result, both groups may support the initiative, but still have a very unrealistic idea of what it involves. Too often these groups have the attitude that it's just metadata – so when they discover that it can't be done in a couple of weeks, or with some new programming technique or even right out-of-the-box, their support evaporates.

Because this is an unfamiliar area for many, it is up to the advocates to educate, communicate with and win support from both business and IT stakeholders. One danger to note in relation to educating management about the complexity and scope of enterprise content categorization is this: it's easy to give the impression that this effort is so esoteric and complex that it is dangerous to attempt and may be beyond their comprehension. Avoid making this mistake with management.

Project mindset: this is a regular software project and is "done"

Most enterprise content categorization development does not follow a typical project model that requires a major effort for a relatively short time period, and then is complete after achievement of a predefined milestone. The danger of this "incorrect mindset" often seems to be understood initially, but many still fall prey to traditional business and IT biases.

Further, just as enterprise search efforts have failed because of a lack of ongoing resources, enterprise content categorization initiatives are vulnerable to the mindset that once the project is done you don't need to allocate resources to maintenance or governance. But unless you are a small organization, assigning one part-time resource to keep your enterprise search effort going – much less growing – is simply not enough. And for enterprise content categorization, there is even more need for ongoing resources – with ongoing taxonomy maintenance and upgrades, ongoing application integration and so on.

All the same it is important to avoid over-stating the need for ongoing resources. If the system is well designed, you can often get most of your needed resources from part-time contributions.

Not enough research – and talking to the wrong people

Another problem that often shows up in the initial development phase is not doing enough research into user information behaviors and needs, and how those behaviors and needs fit into business processes. How much is enough is,

of course, hard to generalize. But one thing is certain – you have to talk to users rather than relying on what developers and managers think their users want. And talking to users does not mean simply asking them what they want – because the answer is always: "all of the above."

Rather than asking what they want in terms of information access and relevance, a good researcher asks the behavioral question. That is, good researchers ask what they do and how they get the information they need within specific activities – then, they build requirements on that basis. Another thing I have found is while it is important to do actual interviews with selected users, the selection process usually still misses something. So it is important to cap your research and your preliminary findings with a broad survey, to test those findings and uncover the missed requirements.

Not enough resources, wrong resources

Because of the complexity and scope of enterprise content categorization initiatives, these projects are often under-funded and resources are not given the time necessary to do a really good job. Again, this underscores the need for good communication to get all the stakeholders to understand the full scope of the project.

Also, in many organizations, most resources go into programming and project management. So typically, there are not enough resources on the semantic side. And since this is a new side, and one that is often not fully understood, it is easy to overlook the need for additional resources.

Enthusiastic access to content and people

As a consultant, I'm often told that I should not bother users whose time is too valuable to spend on interviews. This is one of those hidden dangers that is hard to overcome, especially if the organization has already tried two or three search or information access projects that failed to deliver. In this case, the real fear managers have is that users will see this new effort as a colossal waste of time.

While good communication can help get management on board with project plans, it is really up to management to communicate to their employees how this will ultimately give them more time. It is also important to get enthusiastic access. It does little good if – like one interview I still vividly remember – the person being interviewed is participating under duress, giving monosyllabic answers throughout the interview while doing e-mail at the same time.

Another area that can doom enterprise content categorization is getting access to the content. This might seem like a no-brainer, but it definitely happens – either because of security fears and restrictions, or due to simple technical issues. Using a few hundred or even a few thousand selected documents to build a categorization taxonomy that will be tasked with categorizing millions of all kinds of documents simply will not work.

Bad design: starts with a bad taxonomy

Not all taxonomies are equal. Some are good, some are bad and some are downright grotesque. And just as importantly, some might be fine for their original purpose (say, browsing to find information on the corporate intranet) – but they may not be well suited to the kind of taxonomies that work best for categorization.

Some issues are associated with standard taxonomy design, like having too many top level categories or having chaotic and hard-to-understand category relationships. But some issues are specific to categorization, such as having well-defined characteristics to distinguish overlapping categories all at the same level, like Information Management, Information Technology and Information Security. These design flaws are bad enough with conventional taxonomies, but with categorization taxonomies the flaws make it virtually impossible to develop good categorization rules.

Categorization is more about cognitive science than library science.

Categorization is not library science

One last danger worth mentioning relates to librarians who become categorization taxonomists. While librarians can make very good categorization taxonomists, it is important to recognize the differences between library science and categorization, and the subsequent need for additional training to turn librarians into categorization taxonomists.

The danger of simply having traditional librarians try to do categorization is that they will produce elegant and complex systems that work well – but they will only work well if you are another librarian. The categorizations librarians tend to create don't usually support users well in finding information for their jobs.

Categorization is more about cognitive science than about library science. It is about developing systems that reflect how people use their inherent categorization capabilities – how they perceive, mentally organize and evaluate things to make sense of the world in ever-changing ways. It is not some huge, formal system with a place for everything and with everything staying in its place.

Final Thoughts

Now, all this might seem absolutely overwhelming in its complexity and scope and cost – and it would be, if you had to do everything all at once while learning entirely new fields. However, it is important to remember a few things. First, it is not as complex and costly as it might seem to someone in IT or on the business side of the company who has not had any experience with semantics and taxonomies. Developing content categorization capabilities and applications can appear as a mysterious black box, but for those of us who have been working in this area it is actually less complex than typical programming applications. In addition, the cost is typically less than a single information application – yet a semantic infrastructure component is something that can be used in multiple information applications. Second, the value you will get from content categorization will be much, much greater than it seems. (For more information, see the companion paper *Enterprise Content Categorization – The Business Strategy for a Semantic Infrastructure*.) And finally, there are proven techniques – such as the ones detailed in this paper – that we know work and that will deliver real value to organizations in every industry.

The best way to approach an enterprise content categorization initiative is to think big, start small, scale fast.

It is also important to remember that enterprise content categorization initiatives should always start by developing a complete understanding of the semantic infrastructure of your organization. That understanding will be your foundation for success. It is also important to remember that developing good enterprise content categorization requires a significant human effort, so you must beware vendors who promise "automatic" solutions.

To continuously deliver value to the organization, successful enterprise content categorization requires an infrastructure approach – not a project mindset. However, even though a successful enterprise content categorization initiative will ultimately transform the way information is accessed – in every corner of your enterprise and for every employee – it does not require you to first build a giant, enterprisewide solution before it begins to deliver benefits. If your foundation is in place, you can start with an applied, small project that serves as a learning tool. With this approach, you will start to see immediate benefits. And once you have one project built on that foundation, you can quickly and easily build additional projects, using the combination of a strong foundation and the learning experience gained in the first project.