



▼ Applied  
▼ Relevance

## Agile Taxonomies

A Sensible Approach to Organizing, Tagging and Searching Unstructured Information in the Enterprise.

# Introduction

Taxonomies are a natural way to organize information. The human brain thinks in taxonomies and we can leverage that tendency to make it easier to find enterprise information.

Over 80% of all enterprise information is unstructured. Organizing and finding it is a challenge. The good news is we can use our natural instinct for hierarchies to organize information and find it later.

This document describes an agile approach to organizing unstructured enterprise content. Once you have added structure to unstructured data, it is much easier to navigate and search.



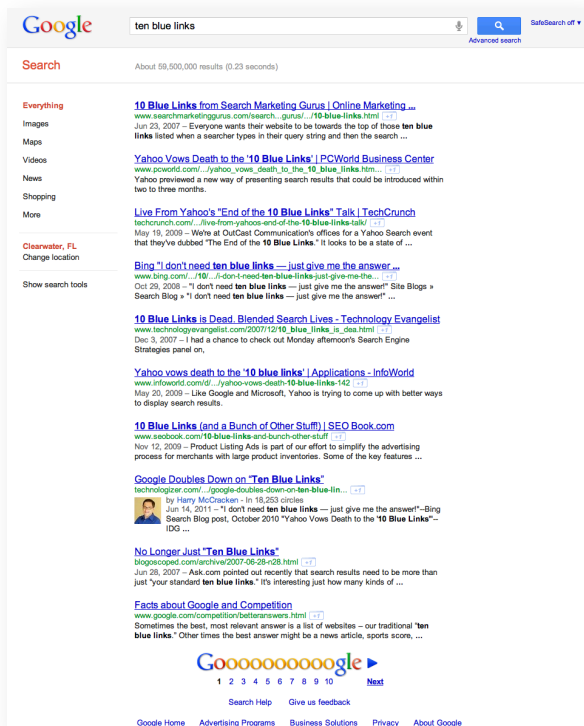
## The Business Problem

Here's a shocker: the World Wide Web is bursting with information. There are billions of pages and links covering hundreds of subject areas that can be found with a simple keyword search on Google or Bing.

What a marvelous time we live in!

Inside an organization is a different story. Even though there is much less information, it is much harder to find.

Why is this? Why is it so easy to



find information on the World Wide Web, and so difficult to find it inside your own organization?

## Ten Blue Links

Search relevance is all about signals. The top pages in a typical "ten blue links" search result list is defined by literally dozens of pieces of information about the information. Metadata, if you will.

Larry Page of Google famously invented the "Page Rank" algorithm. This constantly changing formula looks at features such as "in-links" and "out-links" as well as page title, meta tags and URL to determine what pages best match a given keyword.

Microsoft takes a similar approach for it's ranking algorithm in Bing.

I think we can all agree that Internet search is pretty darn good, and has been for some time.

## Mixed Signals

If ten blue links works so well for the web, why doesn't it work in the enterprise?

The answer is signals. Inside the corporate firewall we just don't have the same signals that are available on the Web. What's different then?

For starters, most documents are isolated. There are no "in-links" and "out-links" to mine. Next, most documents are in Office or e-mail formats, not HTML. There are silos of information with access restrictions, maybe locked-up in a proprietary document management system or database repository. Finally, even the largest organization has a mere fraction of the total users on the World Wide Web, making

click related tracking nearly useless.

This does not mean that there are no signals in enterprise documents. It simply means that the signals are different.

## Signal Generators

What kind of signals are there in a typical enterprise content repository? Well - there's the file

*Why isn't my enterprise search as good as a Google or Bing search on the Internet?*

name. And maybe a folder structure. If you're lucky there is a meaningful title. Even luckier and there is good Metadata associated with documents - but who are we kidding? While the Web is a world wide tapestry of metadata linking information together, the enterprise is an isolated Siberian gulag where documents live solitary, meaningless lives with nary an interaction outside of their folder cells.

Good metadata makes it easier to find documents. You can use the metadata for refining search results and for navigating and drilling down to the right answer in few mouse clicks.

Good metadata makes good search results. Great! Let's add some metadata! Who's with me?

Anybody?

Bueller?

Right.

Nobody likes to manually tag documents with metadata. It's like paying taxes or getting a flu shot. It will have some benefit down the road to somebody, but for instant gratification it leaves a lot to be desired.

So people don't do it.

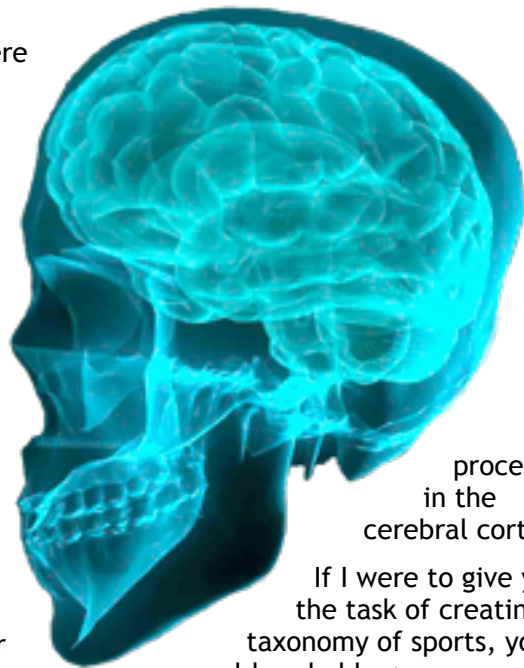
Even if you force them to do it, they do it wrong and resent you for it.

---

## Taxonomies as Signals

---

A taxonomy is a hierarchy. It is a natural way to think about the world around us. You are familiar with taxonomies whether you know it or not. The human brain is physically, not just metaphorically, structured using hierarchical



processing  
in the  
cerebral cortex.

If I were to give you the task of creating a taxonomy of sports, you could probably go away and come back in a couple of hours with a pretty convincing taxonomy of sports.

You and I may take slightly different approaches, but just about anybody could look at what you created and say, "That's about sports!"

I might start with types of sports: team sports, individual sports, Olympic sports, etc.

You might start with regions and teams. But in the end, what we came up with would be much more convincing that even the most advanced "concept extraction" algorithms in existence today.

---

## What does it mean?

---

There is a whole branch of computer science dedicated to extracting "meaning" from text. It is closely aligned with artificial intelligence in the loftiness of its goals. These sets of linguistic and statistical techniques have no less ambitious a goal than allow computers to "understand" ideas so that they can be used to augment human intelligence. It is indeed a

*Nobody likes to manually tag documents. It's like paying taxes or getting a flu shot.*

noble, worthy pursuit. Some of the smartest people in the world are working on this problem - and are dedicating their lives doing so that some time, HAL 9000 will not be a fairy tale.

Taxonomies by themselves are quite simple. There is a hierarchy of parent-child relationships among the terms. We all understand it because taxonomies are intrinsic to human cognition. Like our “sports” analogy above, If I were to give you a topic, say, “Music” I’ll bet that you could come up with a decent taxonomy. I could, too. Our taxonomies would not be identical, but I bet I would understand yours and you would understand mine. They would make sense.

Now, let’s give that task to a computer. “Hey, computer, what’s a good taxonomy for music?”. A common way to do this would be to feed the computer a million documents about music. We call this “training” the computer. It would go away and do some statistical computations on the frequency of words and phrases and maybe do some noun phrase detection and natural language processing and come up with some kind of “hierarchy” of “concepts”.

The problem is, it would be rubbish. At least it would be rubbish compared to the two taxonomies that you and I popped off the top of our heads in a few minutes of human thought. That’s because the human brain is a very high-quality pattern-matching engine that thinks in taxonomies.

Computers are much more literal. They are powerful tools, but asking a computer to build a taxonomy is like asking a computer to build a house.

Computers will be able to build a decent taxonomy when computers can fall in love. It may happen.

But it’s been five years away for the past 60 years, so don’t hold your breath.

The smartest algorithms in the world do not understand meaning. Let me repeat that. Computers do not understand meaning.

Does that mean that all of these fancy linguistic analysis, semantic processing, text mining, information retrieval techniques are useless? Of course not! All of these disciplines certainly have their place and can do many very interesting things. It’s just that they can’t build a decent taxonomy.

---

## The Agile Approach

---

The agile approach to taxonomies, auto-tagging and information retrieval is simple. The human brain uses hierarchies to model the world. We can exploit that property of cognition to help people find the information they need.

Maybe we should state the business problems. Countless studies and surveys have been conducted that identify the following problem areas for enterprise information access:

- ▶ People can’t find the information they need to do their jobs.
- ▶ Documents are not well organized.
- ▶ People hate to manually tag documents.
- ▶ Documents are on one side of the room, taxonomies are on the other.
- ▶ Managing categories and ontologies is complex.
- ▶ Computer generated taxonomies are awful.

While solving this problem, we’ve come up with the following core principals:

*The smartest algorithms in the world do not understand meaning. Let me repeat that. Computers do not understand meaning.*

## Simplicity

Any system must be simple to use. I must be able to describe it to my Mom, who is a smart woman, but not technically savvy.

## Determinism

It must be deterministic. That means that you need to be able to

## Metadata Magic

Any librarian or knowledge engineer will tell you that Metadata is key. Metadata is information about the information. File names, dates, content type, language, author - these are all typical Metadata. Metadata is one way to find what

### The Agile Manifesto

We are uncovering better ways of developing software by doing it and helping others do it. Through this work we have come to value:

*Individuals and interactions over processes and tools*

*Working software over comprehensive documentation*

*Customer collaboration over contract negotiation*

*Responding to change over following a plan*

That is, while there is value in the items on the right, we value the items on the left more.

figure out what's going on in the engine so that you can have confidence in the results.

## Naturalness

Naturalness means that the solution must not feel contrived. It must leverage our natural propensity to organize the world in taxonomies.

## Getting the most out of Agile Taxonomies

In modern software engineering, a project methodology called "Agile Development" is quickly gaining prominence as the preferred method for building software.

In building our taxonomy management software, we try to adhere to these principals. They have proven to be quite effective.

Since it has been so effective, we thought to ourselves "Agile techniques are great for building software. Could they work for building taxonomies as well?"

you are looking for. For Example:

*Show me all documents by Bill Buzby written between June 1, 2009 and June 30, 2009 that contain the word "hoopla".*

That's Metadata plus a full text query. All enterprise search engines can do this, no problem.

Going back to the Web - there's all kind of Metadata either put there on purpose by search engine optimization people, or which can be inferred from the structure of the Web itself.

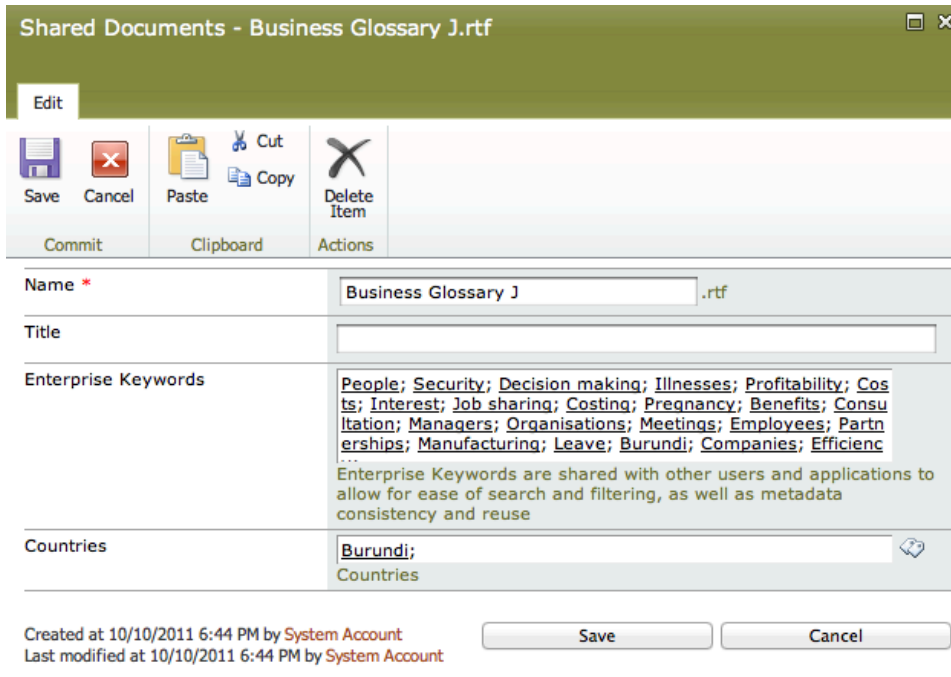
There is no such Metadata available in the enterprise.

If you just spent 40 hours writing a masterpiece of technical literature, there is very little incentive to even add a coherent title and the actual author name to the document. It is inconvenient.

## Active Taxonomies

What are "active taxonomies"? Simply put, an active taxonomy is

"Computer Science is a Liberal Art" - Steve Jobs



one that does something. Traditionally, taxonomies exist for their own sake. They are sitting on one side of the room as a data model for unstructured information. On the other side of the room are documents; often multitudes of documents.

Active taxonomies have a tagging rule associated with each node in the taxonomy. This “rule” term is how the document will be tagged with taxonomy information. This is a powerful technique that is easy to use and understand.

- ▶ Create or import a taxonomy
- ▶ Add rules to the elements of the taxonomy
- ▶ Index the documents against the rules
- ▶ Use resulting facets to search your documents

## Automatic Tagging

“Tagging” a document is the act of adding metadata to the document. It is often called “content enrichment” in that it is adding

information to the document that makes it easier to find.

We mustn’t forget our end goal here - to make it easier to find information in the enterprise. Tagging can certainly help with that, whether it is manual or automated.

The way automatic tagging works is as follows:

- ▶ A document is added to the system. This can be a check-in to a content management system or a crawl from a content repository.
- ▶ The document is passed to the “Annotator”, which matches all the taxonomy rules against the full-text of the document and it’s metadata.
- ▶ Any matching terms are sent back to the calling system, which adds them to the document metadata properties.

The beauty of this workflow is that nobody had to manually add metadata to the document. It got populated automatically.

The real “a-ha” moment comes when the taxonomy terms are

*The real “a-ha” moment is when the taxonomy terms are automatically added to a document at check-in - instantly.*

automatically added to the document at check-in. It's magical.

## Faceted Navigation

The final piece of the puzzle is the search interface itself. How do you leverage the taxonomies and tagging that we've created so far?

Luckily, most modern search engines have a feature called "Faceted Navigation" or "Parametric Search". This feature allows us to create relational taxonomies that interact to let you drill down to exactly what you are looking for.

Relational taxonomies help you find what you know and discover what you don't know.

If you've ever bought anything online, you've used faceted navigation. They are the lists of product features, usually in a left navigation window that are dynamically filled with values.

Say you're looking for a car. On the left, you may see make, model, color, year, each with lists and numbers next to them.

- ▶ Audi (20)
- ▶ BMW (2)
- ▶ Cadillac (132)
- ▶ Datsun (3)
- ▶ Rambler (1020)

This example would be the "make" facet.

You can click on the items and drill down to the exact car you are looking for, very quickly.

That's where the power of Active Taxonomies kicks in.

## Conclusion

Active taxonomies help end users find what they know and discover what they don't know.

Automatic tagging enriches document metadata with the

structure that is buried in unstructured content.

Faceted search lets you drill down on the precise information you need with just a few clicks.

The screenshot shows a search interface for 'soda' with a left-hand navigation menu and a main results area. The left menu is organized into categories like 'General Business', 'Food and Beverages', 'File Type', 'Modified By', and 'Content Type'. Each category has a list of sub-items with counts and checkboxes. For example, under 'General Business', 'Advertising' has 216 items, and 'Marketing' has 204. Under 'Food and Beverages', 'Food Services' has 216 items, and 'Beverages Services' has 216. The main results area displays a list of search results, each with a title, a snippet, and metadata like 'Modified by' and 'File Size'. The results are numbered 1 through 10, and there are navigation buttons for 'Prev' and 'Next' at the bottom.

# ▼ Applied ▼ Relevance

Applied Relevance, Inc.  
Saint Pete Beach, Florida USA  
+1 727-498-0222



[www.appliedrelevance.com](http://www.appliedrelevance.com)  
[info@appliedrelevance.com](mailto:info@appliedrelevance.com)



[linkedin.com/company/applied-relevance](https://www.linkedin.com/company/applied-relevance)



[apprelevance](https://twitter.com/apprelevance)



[facebook.com/apprelevance](https://www.facebook.com/apprelevance)



[youtube.com/appliedrelevance](https://www.youtube.com/appliedrelevance)