Folksonomy, Keywords, & Tags: Social & Democratic User Interaction in Enterprise Content Management

*An Oracle Business & Technology White Paper*
*July 2007*
INTRODUCTION & OVERVIEW

Information, content, is being created at an ever increasing rate. The information explosion is fueled by three converging trends that span technological, cultural, and regulatory disciplines.

1. Technological: the proliferation of technologies that facilitate content creation. Examples include blogs, wikis, social network sites like LinkedIn, MySpace, social video and image sites such as YouTube and Flickr.

2. Cultural: the emergent web 2.0 mindset that promotes, with a force approaching the demands of a human right, hyper-individualized and perspective-based content creation.

3. Regulatory: Legal statues, business regimes and best practices which increasingly require the persistence of nearly all information created. Whether for risk management, corporate compliance, or from a desire to provide better metrics through richer data sets, information is kept longer than ever before.

The result of more people more easily creating more content which is then stored for more time is info-glut: an overabundance of information. Businesses are catching on to the need to manage this information with enterprise wide content management strategies that provide a unified foundation or infrastructure which facilitates the management of content, in all its disparate formats (e.g. documents, images, videos, websites, API-messages etc.), and allows for the securing of content even after it exits the enterprise content management (ECM) systems.

Yet more content than ever before, even when properly managed in an ECM system, is still more content than ever before. The challenge businesses face today is not an inability to create content, or even an inability to manage content. Rather it is a three-fold challenge:

1. to provide the right content
2. to the right people
3. at the right time

This is the challenge of relevancy.
Info-glut obfuscates relevancy. Rather than challenges from previous eras where creating the right content was difficult, info-glut tantalizes us with the knowledge that relevant information is in there...somewhere. Info-glut is the haystack atop the needle. It is the sea in which the treasure resides, a tsunami for want of a cup of water.

**KEY BUSINESS REQUIREMENTS**

At its core, the problem of relevancy is a search result problem. The question of relevancy presupposes a “to what” predicate. A search result can only be relevant (or not) to what the searcher wants. Only secondarily is the search result relevant (or not) to a search term, query string, or criteria. This means that the possibility of relevancy is a function of how well a search term actually reflects the searcher’s desire. The ability of the searcher to accurately translate his/her desire into an appropriate middle language (query string, form value, criteria, etc.) is an implicit limiter on the degree of relevancy that is possible in the search results. For example, if a user desires information on Australian Shepard Dogs and is able to only search for “animal”, the relevancy of the results is substantially diluted.

As such, many of the early attempts to deal with relevancy sprang from a structured content mindset: database thinking. Search forms were more user-friendly ways of writing database queries than SQL. However, these only dealt with relevancy indirectly. Any and all results that came back were assumed to be relevant simply because they matched the select query criteria. In an info-glutted system, criteria-matching is no longer enough to assure relevancy. When searches yield hundreds or thousands of results the relevant item is effectively buried, and therefore, lost. People who find exactly what they’re looking for at the top of a results list should immediately buy a lottery ticket; they’re in a luck-warp.

Fortunately, different approaches have arisen, organically, from the emergent web 2.0 world to deal with the problem of relevancy. These approaches succeed or fail to a greater or lesser extent. But one theme that many of these strategies share is a social or democratic approach to the solution. Because relevancy is determined by the seeking audience, when that audience helps to define relevancy, better relevancy scores or rankings are the result the next time around. The idea is that the more what it is, and how it is used information the seeking audience affiliates with the content, the greater the accuracy of the relevancy algorithms is the next time they are applied. But relevancy is also expressed in the collective agreement of groups. This is the power of social and democratic approaches to relevancy solutions: fostering the hyper-personal interaction capability for inputs and watching amalgamations that heretofore were not conceived precipitate out (or percolate up) from the interactions of the individual contributions. This is the power of the social and democratic approaches to relevancy. Four advantages to a social or democratic approach to determining or increasing relevancy are immediately evident.
1. **Decreased central burden.** The social approach is decentralized. It relieves the burden on and responsibility of a top-down or centralized approach. Centralized approaches such as corporate taxonomies, pre-defined option lists, normalized database schemas, and myopic *most-recent-therefore-most-relevant* approaches place an enormous burden on human staff such as central administrators, corporate librarians, records managers, as well as on technologies such as search indices and query engines. The cycle of tuning and tweaking the systems in order to better serve the end-users can become never-ending.

2. **Increased accuracy and granularity** of relevancy. Relevance is directly proportional to specificity. A social approach to a relevancy solution tends toward greater accuracy and granularity since it takes a bottom up approach to relevancy determination. Rather than relying on a central authority to establish relevance across all the permutations of ranking and ordering of results to all the permutations of queries that the end user community might use, the social approach allows the user community to establish relevance on the go, as they use the content they find. The mechanisms by which this can and should happen are discussed below.

3. **Self correcting, self tuning results.** A social approach to relevancy is inherently self correcting and self tuning. As social and democratic input is added over time, anomalies are statistically marginalized to the point of immateriality and gaps are filled in. In this manner, a system learns and corrects and then provides the ever increasing levels of accuracy and granularity which determine relevance.

4. **Hip to Web 2.0 culture.** A democratically based solution to the relevancy problem is inherently web 2.0. It invites and requires participation by end users. It is radically decentralized. It fosters reuse and repurposing of content yielding efficiency gains. It reveals a rhizome-like potential for content that often transcends original intent and represents a nascent semantic web.

But what exactly *is* the social or democratic approach to a relevancy solution? Several concrete examples from the World Wide Web should help describe the various socially based solutions to relevancy.

---

**Google™**

The Google™ search engine measures inbound links (among other factors) to the websites it indexes to determine relevancy. The more inbound links (i.e. other sites that link to it) a site has, the higher it’s ranking. The implicit presumption is that links to a site act, in part, as a currency of relevance. Inbound links are an arbiter of one of the few items of value in the online world: credibility. The “richer” the site,
the better the chances it will provide relevant content. Ergo, it gets listed higher up in the search results. To be sure this is an oversimplification and there are other factors, such as purchased keywords, query matching against result candidates, etc. that can affect ranking display (for a Google insider’s perspective on this see David Elworthy’s blog post from March 15, 2006 here). Nevertheless, the point stands. After all, other web site owners would not link to sites without “cred”, would they? It should be noted that there are some mischievous individuals who seek to manipulate the perception of credibility through popularity analysis. These attempts are real though they prove the inherent point: popularity is a presumptive indicator of credibility. A more detailed discussion of popularity manipulation through such phenomena as “Google bombing” is discussed below.

For determining hipness or popularity such a model is fine. This approach, however, does not so much determine relevance as much as it leverages a [potential] side effect of likely relevance. Such a possibility stems from the expectation that a popular site will say something important on the topic. For determining relevance, especially in relation to the relevance of information contained in lesser ranked sites (read: sites appearing in page 2+ of the Google™ search results), this approach provides little to no relevancy benefit. In other words, a site appearing on page 9 of a Google™ results list may be the best site “out there” despite the fact that it has few inbound links.

Despite the shortcomings of the Google™ model, it does serve to exemplify the social or democratic approach to relevancy and, at the end of the day, it does a pretty good job of pointing us toward the information we seek.

**Flickr™**

The photo sharing site Flickr™ is an example of a site that first popularized the tagging concept. Tagging on Flickr™ is, quite simply, a form input field whereby the photo owner may apply one or more keywords, or tag, to a photo that s/he has uploaded. Images are notoriously difficult to index in meaningful ways due to the lack of machine-readable, human-understandable, textual data associated with them (as opposed to file size or extension for example). The quickest way around this snag is to simply associate human-input textual data with the image and then leverage that data, in aggregate, as a ready-made index against which searches may execute. Flickr™ implemented this notion with their photo tagging capability. They then took this one step further and aggregated the most popular tags into a tag cloud; a visual representation of the tag aggregate in which more common tags are differentiated from less common tags by font size and font weight.
The tags in a tag cloud represent subsets of content that precipitate organically out of the aggregate or totality of data. Users with no connection to each other may tag their content with the same word and thereby, independently, democratically, add it to the (or define a new) set of content that then becomes available for the end-user community. Such an organic and subversive (to the search engines and indexers at least) approach is immediately and intuitively appealing to a web 2.0 mindset.

Tag clouds, too, are incredibly powerful mechanisms and substitutes/alternates for form-based query builder search tools. By presenting the user with a visually represented search interface, the cloud taps into the power of the human brain to understand, inherently, what it wants better than a query language can translate that desire into a machine-understandable request. By aggregating and presenting keyword tags, the user is invited to drill down into the subset that most likely contains the content in which s/he is interested. In the Flickr™ example, larger tags indicate larger quantities of likely data in that subset while smaller tags indicate smaller subsets of data.

However, while tagging is an individualistic activity, relevancy can still remain an elusive goal. There is no suggestion engine or central application engine that enforces tags on images. Subsets are defined democratically insofar as individuals make their own decisions on what tags to apply to their content. Yet, if one user tags a photo of his Australian Shepard dog with “dog” and another user tags a photo of her car with “dog” both photos will display in the set of photos containing the “dog” tag. The results are only as good as the fidelity of the tag to the image to the end-user’s expectation. Maybe the woman’s car is a “dog of a car” or maybe she calls it her “dog”. Nobody knows except the tagger and possibly those in her immediate circle of friends. These colloquial or micro-social contexts are neither known by the end-user community nor communicated by the tag “dog”. Therein lies the challenge that owner-defined tags pose to relevancy. Social tagging relying only on content owners for input is beholden to the whim of the owner and whatever taxonomic baggage she carries with her.
Despite the shortcomings of the Flickr™ owner-tagging model, it does serve to exemplify a social or democratic approach to relevancy and, at the end of the day, it does a pretty good job of pointing us toward a set of information that probably contains what we seek.

**Last.fm™**

There is another kind of social tagging that presents in clouds similar to the Flickr™ model but that takes a very different approach to how tags are created. Instead of content owners being solely responsible for applying tags to “their” content, the entire end-user community is invited to tag content. In this manner the audience defines the content. As participation by the audience increases (as measured by the number of tags applied over time), certain terms emerge as more common while others remain outliers. In this way, the end-users “vote” on the term or terms that best describe or “fit” the content. Over time the “fit” gets better as more “votes” are tallied.

The music site Last.fm™ uses this style of democratic tagging to label the genres of the artists they host. The cloud that is formed from such an audience-based tagging approach typically indicates a single leading tag. But rather than declare a artist’s genre definition as the one “winning” tag, the entire cloud is left such that the relation of the “winning” tag to the other tags is apparent. This has the effect of “flavoring” or moderating the genre. Consider this cloud for the band “Arctic Monkeys” appearing on the Last.fm™ site:

While the presentation is identical to other tag clouds, what is interesting is what the cloud as a whole communicates to the end user and the implications on relevancy. This artist is clearly tagged “indie” or “indie rock” – arguably two synonymous or nearly synonymous genres. But “alternative” and “britpop” and “singer-songwriter” also appear in the cloud. From this the audience can generate a much richer inference about this artist, their likely sound and influences than simply “indie rock”. The cloud, itself, is a living metadata asset to the content – in this case the artist’s page on Last.fm™.

The cloud serves another purpose as well. This cloud, as many clouds are, is a collection of hyperlinks. Clicking on “indie rock” allows the end user to drill down into a collection of other artists who have been similarly

---

**Top Artists tagged “indie rock”**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Artist</th>
<th>Tag Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Death Cab for Cutie</td>
<td>2,291</td>
</tr>
<tr>
<td>2</td>
<td>The Arcade Fire</td>
<td>2,114</td>
</tr>
<tr>
<td>3</td>
<td>Billy Joel</td>
<td>2,003</td>
</tr>
<tr>
<td>4</td>
<td>Franz Ferdinand</td>
<td>1,783</td>
</tr>
<tr>
<td>5</td>
<td>The Breeders</td>
<td>1,732</td>
</tr>
<tr>
<td>6</td>
<td>Arctic Monkeys</td>
<td>1,712</td>
</tr>
<tr>
<td>7</td>
<td>The Killers</td>
<td>1,593</td>
</tr>
<tr>
<td>8</td>
<td>The Glories</td>
<td>1,444</td>
</tr>
<tr>
<td>9</td>
<td>The Shins</td>
<td>1,421</td>
</tr>
<tr>
<td>10</td>
<td>Pavement</td>
<td>1,401</td>
</tr>
<tr>
<td>11</td>
<td>The Decemberists</td>
<td>1,154</td>
</tr>
<tr>
<td>12</td>
<td>The White Stripes</td>
<td>1,132</td>
</tr>
<tr>
<td>13</td>
<td>Yeah Yeah Yeahs</td>
<td>1,101</td>
</tr>
<tr>
<td>14</td>
<td>Broken Social Scene</td>
<td>1,084</td>
</tr>
<tr>
<td>15</td>
<td>Fables</td>
<td>1,001</td>
</tr>
<tr>
<td>16</td>
<td>Bright Eyes</td>
<td>953</td>
</tr>
<tr>
<td>17</td>
<td>Kaiser Chiefs</td>
<td>955</td>
</tr>
<tr>
<td>18</td>
<td>Snow Patrol</td>
<td>951</td>
</tr>
<tr>
<td>19</td>
<td>Neutral Milk Hotel</td>
<td>915</td>
</tr>
</tbody>
</table>
tagged by the audience of Last.fm™ users and listeners. From a search relevancy perspective, this has the effect of providing a list of results relevant to “indie rock” and related or similar to the “Arctic Monkeys”. For the end user, such a related relevancy engine exposes him to other content (artists) with whom he may not be familiar or have considered and provides a visual indicator of how accurate the tag (“indie rock”) is as a metadata indicator (in this case, genre).

Gaps, Deficiencies, & Missing Features

The three case studies above illustrate the general state of democratic approaches to the relevancy issue. To be sure, there are sites and developers who have implemented relevancy strategies, some of them socially or democratically oriented, which surpass the strategies of Google™, Flickr™, or Last.fm™. There are also those who have not started to consider a relevancy strategy, whether they are owners of public facing websites or administrators of internal networks. But before organizations begin to consider a relevancy solution, be they technologies, development approaches, or off the shelf applications, they should consider what is missing in, but needed from, the current approaches.

Missing from popularity analysis of inbound hyperlinks is an evaluation of inherent as opposed to probable relevancy. Missing from author-originator keyword labeling is an ability to transcend the ontological (i.e. what *is* or *is contained in* the content) baggage and teleological (i.e. why or what *is the purpose of* the content) preconceptions about the content that are imported by the author. Missing from the democratic or “voter” style of tagging is both any kind of data seeding as well as any ability to attribute tags that capture the ontology and original teleology of the content. These concepts are explained below. Finally, missing from both of the previous two “tagging” approaches is an understanding and appreciation of the differences between and power of owner-originator keyword aggregation and social tagging.

A brief digression on the distinctions between content ontologies / teleologies, the distinctions between democratizing / socializing technologies, and the distinctions between keywords / tags may be in order.

With the risk of oversimplification, a content ontology is a way of describing the concepts in a piece of content (which a metadata model does) as well as the relation of those concepts to other concepts in the content and other content items (which a metadata model does not do). Simplistically, content ontology is *what the content is*. The semantic web project and various standards bodies are very focused on enabling and defining content ontologies and the structures that support them (e.g. RDF, OWL) for the purposes of allowing machines to make ontological inferences and thereby provide users with more/additional relevant, though perhaps initially unrealized/un-asked-for content. A content teleology is a way of describing the purposes of/in/for/about a piece of content as well as the relation of those purposes to other purposes for that and other content items. Simplistically, content teleology is *what the content is for*. While these two concepts may be easily conflated
as “content ontology” they are, in fact, different and distinct and, when identified and leveraged distinctively, hold great potential for answering and solving or at least mitigating the relevancy problems that infoglut poses, if not in the vast and wild expanses of the World Wide Web’s universe then certainly in our more tamed and manageable back yards of our content management systems.

Democratizing and socializing technologies are both accessible to the people, having little to no barriers to entry and participation. However, socializing technology is understood in this context to be accessible to the people as well as allowing/fostering/requiring a synthesis, or at least colloidal blending, of concepts. It is a multilectic process where n-inputs (e.g. tags or keywords) are aggregated to impact upon each other and form a new whole (e.g. a cloud). This new whole is a new kind of object that contains traces of the constituent parts but where no constituent part goes un-flavored or un-nuanced by the others. Dissimilarly, democratizing technology, while accessible to the people, is understood in this context as allowing/fostering/requiring a tallying of inputs. Inputs remain distinct. If aggregated, the aggregate is a report or a display rather than a uniquely new whole. The aggregate provides a presentation of multiple tallies for purposes of visually determining ranking, preference, or popularity.

Tags and keywords are also differentiated along a similar binary opposition. In this context tags are understood to be inherently connotative, social and possibly democratic while keywords are understood to be inherently denotative and possibly democratic but rarely social. Keywords are supplied by author-originators and are denotative. This means that they are provided by a small set of individuals (usually one or two) with the intent of standing for or being signs of. As such, they are the carriers of the seeds of an organic content ontology but are not, in themselves, a content ontology. They contribute to and jumpstart a formal content ontology that is fleshed out with other formal metadata structures and possibly RDF and OWL attributes. Tags, on the other hand, are supplied by consumers (who may include but are not limited to author-originators) and are connotative. This means that they are provided with the intent of signifying or suggesting concepts that are potentially accompanying or associated with possible content ontologies and teleologies. In this sense individual keywords represent a ‘what the content might be or mean, to whom’ while tags represent a ‘how or for what the content might mean’.

From an implementation and application perspective, the ways in which keywords and tags are used is also best differentiated. Keywords, as author-originator selected (more on this below), are ideally situated for one time application or once per content revision application and then tallied and presented according to democratic principles. This best taps both original intent and purpose as well as provides a snapshot in time of early content ontology and teleology. This snapshot can then be tallied and algorithmically evaluated to perform more in-depth democratic-perspective-over-time evaluations and presented to display current ontology and teleology vote snapshots. Alternatively, tags are ideally suited for ongoing and unstructured application to content. This best taps the organic and
ongoing ontological and teleological context of the user community. When presented back to the consuming community, the meaning and purpose feedback loop is closed and end users may take that information into consideration when making final, relevancy-based selections. If an item is insufficient, the consumer is invited to participate in the ongoing social tagging in a kind of pay-it-forward relevancy model.

**Popularity**

While it is important to capture popularity of, on a macro-level, web sites, or on a micro-level, pages, or on a nano-level, documents, sections and snippets, usage and link tracking are not guarantors of relevancy. Neither are analyses of popularity through inbound links guarantors of credibility. The phenomenon of *Google bombing* illustrates this point. “Google bombing” is the manipulation of the Google™ algorithms to make certain pages appear closer to the top of search results, thereby giving the impression that they are the most relevant. (for more information on “Google bombing” from an insider’s perspective, see the Google Webmaster Central Blog entry from January 25, 2007 [here](#), written by Ryan Moulton and Kendra Carattini). Such tactics and others like them, though, serve to illustrate the central supposition: that popularity, as indicated by incoming links, is presumed to be an inherent indicator of relevancy. Were it not so, Google bombing would not be the kind of problem that demands a public response from Google™. Consequently and at best, popularity analyses are indicators of the likelihood that *something* important (or at least credible) may be found. A relevancy solution should take these metrics into consideration but not stop there. The good news is that these kinds of metrics are relatively simple to track and capture. Web site traffic meters will typically track hits, popularity, and incoming links (or at least referrers). Most business intelligence tools and search appliances incorporate more sophisticated algorithms to track, capture, and report or serve up this information. But a relevancy solution that stops at this point is missing some of the most powerful processors available: the end-users.

**Keywords as Content Ontology**

Content ontology is a concept that is vital to the goal of a truly semantic web. It is also a goal that a top-down metadata model or corporate taxonomy only achieves half-way. As explained above, a content ontology is a way of describing the concepts in a piece of content (which a metadata model does) as well as the relation of those concepts to other concepts in the content and other content items (which a metadata model does not do). When keyword attributes are included in a metadata model they represent aggregates of components of a content ontology. As such they are able to do a decent job of flavoring or providing nuance to the ontology. But they suffer from a myopia inevitably imported from the author-originator. No one can think herself outside her own head. Consequently, author-originator keywords represent a *‘what the content might be or mean, to whom’* rather than a formal specification of what the content is. Insofar as the audience shares the
intent and context of the author-originator, keywords as helper-carriers of a content ontology may be enough to provide a glimpse of what is in there. However, web 2.0 practice, theory, and implementation increasingly realize a divorce between the audience and any shared intent and context with the author-originator. Indeed, one need only think on the proliferation of mash-ups to grasp the scope of this divorce. There is a growing realization of the arbitrariness of author-originator keywords as signifiers of content ontology. While author-originator keywords are a necessary component of a relevancy strategy, they are no longer sufficient to guarantee relevance. What is required is a complementary strategy that encourages author-originator keywords while democratizing and socializing content tagging. Furthermore, author-originator keyword application should be implemented differently from social tagging schemas. To tap the latent power of keywords-as-ontological-signifiers while simultaneously enabling a democratic approach that will rank and display keywords over time two principles are suggested. It should be noted, though, that these principles are not solutions to the problems of the author-originator’s own baggage (as noted above) but rather prerequisites for implementation of this necessary but not sufficient part of a relevancy strategy.

First, keywords should be managed but extensible artifact sets. Second, keywords should be predictable or at least anticipatable by applications. Keywords as managed but extensible artifact sets would solve a critical semantic problem inherent to a free-form keyword input paradigm: the synonym problem. By way of example, consider two synonymous keywords, “401K” and “401-K”. While such keywords are immediately synonymous to human consumers, they are not synonymous to computers. While Thesauri and stemming rules have been implemented with varying success on the computer interpretation side, the user input side remains open to the problem. Without a management constraint, author-originators may variably apply “401K”, “401-K”, and “401 K” as keywords to the same and related content items. A computer system attempting to tally and aggregate keywords for ranking and relevancy purposes will miss the synonyms (or misspellings) and therefore the relevancy signified by the set of similar but not identical keywords. Stemming rules and thesauri lookups cannot keep pace with a social or democratically oriented organic solution and consequently have an overly limited utility (not to mention performance) in such contexts. Keywords applied thusly will contribute to rather than alleviate the problem of infoglut. As managed artifact sets, however, author-originators may be presented with a select list or taxonomy tree of keywords from which to choose. An author-originator is relieved of the burden of having to pick the “right” synonym. The keyword artifacts (lists, or trees for example) should be extensible, however, to allow author-originators to submit new keywords for consideration. Corporate librarians and keepers of the corporate taxonomy would meet regularly to consider, validate, and implement new offerings. In this way, the management of artifact (i.e. keyword) sets is not completely a top-down approach typical of strictly standardized corporate taxonomies. Rather it is a managed approach that enables input from across the
organization and allows the artifact sets to be organic and evolve with the needs and requirements of the enterprise.

This management approach to keyword artifact sets is critical if the artifact sets are to be accessible to machine systems. A keyword list or keyword taxonomy tree (or keyword branch of an enterprise taxonomy) may be exposed as an computer readable object (e.g. a JavaScript object for AJAX manipulation and evaluation). The ability to expose an objectified keyword artifact set as is necessary if computers and code are going to be able to anticipate terms and evaluate keywords in relationship to other keywords as well as their place in a set (presuming that the keyword set is located in something like a hierarchical taxonomy). The more structured an exposed keyword set is, the greater the potential an algorithm has to make relevancy determination and even inferences about the keyword. Relevancy may be determined (in part) by the “distance” of a keyword from a search term in a taxonomic hierarchy. An inference engine (or suggestion engine) is an easy step from a relevancy algorithm as, in its most basic form, it may be little more than polling for other keywords “near” a specified keyword (again in a taxonomic hierarchy) or pulling keyword sets higher up a taxonomic chain. Finally, if keyword artifact sets are managed sets and exposed to/as computer APIs, then convenience factors for end users (author-originators in this case) may be more easily implemented. Type-ahead capability, keyword suggestions, or automatic keyword labeling applications may all leverage such a set. While these capabilities and conveniences may be implemented as stand-alone functionalities, what cannot be missed is that although these are necessary parts of a relevancy strategy, they lack sufficiency as a relevancy strategy. They must be integrated into a larger relevancy discipline.

Tags as Content Teleology

Content teleology is another goal vital to the realization of a semantic web. It is a way of describing the purpose of a piece of content. While corporate taxonomies and application metadata models may allow for a keyword or two related to purpose, there is little flexibility after a purpose attribute has been assigned. Purpose is in the mind and the mash-up of the user. The reusability of content in ways unforeseen and unanticipated by the author-originator is a hallmark of web 2.0. Consequently, the ability to anticipate what a content item may be used for is nigh on impossible. Nevertheless, the ability to catalog and visualize what an item is and was and may be used for holds strong promise. What has before been mistaken as or conflated with content ontology is more appropriately content teleology. For example, tags or metadata attributes such as “press release” or “web site item” or “for print” indicate purpose rather that what the content is. Unfortunately, in an owner-administrator attribution context (i.e. where only content owners or administrator level users are provisioned to add/modify/remove attributes) the ability to keep these kinds of attributes up to date and relevant to now is substantially diminished. Combined with mash-up and other repurposing /
compositing capabilities that are becoming more and more commonplace, these attributes become increasingly incomplete with the passage of time.

Incompleteness risks irrelevancy. The solution is to allow purpose-signifying tags to evolve, to be organic, and to grow from the source of their changing purpose: the consumers. This is an inherently social approach and social tagging enables end users and consumers to indicate both how and possibly how a content item is used. As mentioned above, these capabilities may be implemented as stand-alone functionalities and indeed examples abound. But what is missing are such implementations of purpose capturing tactics integrated into a larger relevancy discipline.

**Author-originator Tags and Keywords as Data Seeds**

Data seeding is a jump-start or springboard set of tags or keywords or both that serve the purpose of preventing initial obscurity. If socially or democratically applied tags are the only data points that may be indexed or that may appear in an end user facing tag cloud then the barriers to inclusion in initial search results are quite high. Consider a new artist added to the Last.fm™ queue. Without either an end-user community “in the know” about the new addition or an alternate way of listing the addition, the new artist will remain in obscurity until someone stumbles upon their listing and decides to be the first to tag them. For any content, obscurity in the infoglut storm is already a significant risk. A tagging solution should not also exact an “obscurity toll” as the price for possible eventual relevancy.

The lack of a communicated original intent is also an inherent problem with purely democratic tagging mechanisms. While it is nearly a guarantee that, in the mashup-loving-socially-empowerd-web 2.0 world, content can be repurposed, the force and thrust of content will still be greatest in its original intent. Content will typically be strongest when consumed according to its original purpose. This kind of *content teleology* should be captured as metadata. While the Flickr™ keywords give a sense of *content ontology*, or what it is about, the lack of a tagged intent or purpose that is able to be aggregated and machine interpreted (ideally) or presented (at least) is a deficiency of both the Flickr™ and Last.fm™ models. The ability to capture the initial thrust of purpose as part of an overarching relevancy approach is still needed. Promotion and enabling of author-originator tags and keywords as a denotative content teleology is one strategy that ought to be integrated into a larger relevancy discipline. Data seeding tags and keywords alone are not enough to guarantee relevancy or even encourage users to participate. Alone they are lacking the structure of keywords, the organic nature of social tags and the power of popularity rankings. They must be integrated into a larger relevancy discipline to realize the potential they hold.
Inherent but Non-Unique Risks in Folksonomy Implementations & Their Solutions

A dash of consideration of the risks and problems with Folksonomies is warranted lest the utopian proclivities of the more excitable new technology advocates take over. First, as with the communication/transmission of all content, the medium is selective in what it allows to pass. So as a tempering influence on anything that smacks of universal/utopianism, it must be recognized that tagging is (currently) an inherently textual process. There is no mechanism for aggregating and rendering image-as-tags, video tags, audio tags, taste, smell or other semiotic carriers that is as immediately accessible to the user as textual tags are currently. Secondly, social and democratic tagging and keyword labeling strategies presuppose user involvement. While democratic keyword application strategies may be configured to compel author-originators to provide textual inputs in a keywords field, neither the quality nor the propriety of the content cannot be compelled. Furthermore, social tagging strategies, by their nature, cannot be compelled. They can be encouraged, users may be enabled, strategies may be made convenient. However, all social input strategies rely on the good-graces of well-intentioned users habituated to provide input over time to succeed. Third, as alluded to above, even when users are actively participating in democratic and social strategies, the information they provide may not be relevant or appropriate. Social strategies will self-correct for this problem over time under the presumption that more users than not will provide “good” information. Democratic strategies that take new end-user inputs under consideration but rely on a core set of administered keyword artifact sets may be effectively inoculated from this problem, but at a sacrifice of flexibility and responsiveness to trends in time. Under social tag structures, the risk of “tag bombing” or flooding a tag input mechanism with distasteful or irrelevant data can severely undermine the goals of enhanced relevancy and user experience. These are three potential risks, undoubtedly among others, that all democratic and social relevancy strategies share. As such they are not uniquely disadvantageous to democratic and social relevancy strategies as such implemented in an enterprise content management system. Rather these problems are indicative of the need for a solution-as-discipline rather than solution-as-commodity. Beware of quick-fix shills that promise comprehensive relevancy solutions if only you buy their newest product.

Key Business Requirements Redux

So far we have demonstrated the following:

1. The overabundance of content precipitated by advances in information and communication technology is masking the content we want when we want it. This is the problem of relevancy.

2. Various solutions to the problem of relevancy implement web 2.0 style tactics. Many of these tactics are democratic or social or both in nature.
3. Democratic approaches focus on tallying and reporting while social approaches focus on aggregating and swapping nuance.

4. The multiplicity of approaches, be they democratic, top-down, social, bottom-up, taxonomic, folksonomic, consumer, or contributor based may be necessary but are not sufficient, in and of themselves, to solve the relevancy problem.

5. What is required is a solution that is a discipline which incorporates best-of-breed enterprise content management technology with enabling democratic, social and legacy approaches to provide the right users with the right content at the right time.

Solution: Oracle Universal Content Management’s Folksonomy Offering

For folksonomy, Oracle Universal Content Management offers a set of enabling technologies and tactics that combine to allow the implementation of a fully realized relevancy discipline. This relevancy discipline is seamlessly integrated with and into, while resting on and presupposing, Oracle’s overarching enterprise content management discipline. Relying on Oracle Universal Content Management’s service oriented architecture (SOA), agile-development-enabling component architecture and the schema API, Oracle’s folksonomy offering instantiates robust social and democratic relevancy strategies that realize tangible relevancy gains and empower end users of all types to contribute. The end result, Oracle Universal Content Management’s folksonomy offering establishes a solution to the relevancy problem.

SOLUTION COMPONENTS

Oracle Universal Content Management and its folksonomy capabilities are a part of Oracle’s overall content management solution. Oracle Content Management provides the three key foundation components of a successful enterprise content management meta-discipline.

1. Oracle Content Management provides a unified strategic infrastructure for content and applications

2. Oracle Content Management can manage all content stores, including legacy systems

3. Oracle Content Management secures information beyond managed environments

Within this meta-discipline, the relevancy sub-discipline resides. This primary - sub relational structure allows the relevancy sub-discipline to tap into and leverage the capabilities of the Oracle Universal Content Management system and the content management context. In short, Oracle Universal Content Management’s component architecture (CA) leverages its SOA and Schema API in order to tap into and extend the metadata framework and schema implementation.
Oracle Universal Content Management’s component architecture is an agile-development enabling capability. It is a development framework that allows and facilitates parallel development of complimentary capabilities that seamlessly hook into the core of the product, Oracle Content Server. A correctly written “component” behaves transparently, as if it were part of the core.

Oracle Universal Content Management’s service oriented architecture (SOA) has been part of the core since its inception. All of the nearly 800 services in the core, and all of the add-on functionality services are able to be called independently via SOAP or JAVA and return java object or XML responses as appropriate. When combined with the Schema API, which allows either dynamic or static publishing of variably relational and variably filtered database views as JavaScript (.js) files, all the constituent pieces of a robust, fully and seamlessly integrated relevancy discipline are in place. All that remains are the implementation details. Some partial examples are proposed below as springboards for further thought and experimentation.

**Democratic Author-Originator Keywords in Oracle Universal Content Management**

Because it is easiest to start with the tactics that have most of their constituent parts already in place, tapping into the inherent flexibility of the Oracle Universal Content Management metadata framework for a democratically oriented keyword solution is considered first. As discussed above, keywords are ideally situated as author-originator provided artifacts. An easy way to implement a simple author-originator keyword system with a keyword cloud for display and prompting for input is described.

The Oracle Universal Content Management metadata framework allows easy creation of new metadata and schema fields. Author-originator keywords are ideally set during contribution. The input field should be large enough to display author-originator inputs and keywords delimiters (commas in the example below) should be automatic or at least commonplace.

| keywords | example, keyword, white paper, democratic, folksonomy |

Additionally as described above, keyword sets should be loosely administered artifacts. Keyword sets should be predefined and presented in preset interface, such as a multi-select list, and also facilitate suggestions from the contributor community. Suggestions should be regularly gathered and considered by an enterprise taxonomy working group and then incorporated back into the user interface. Enterprises should be cautioned, however, about maintaining too tight a grip on taxonomies that are part of democratic relevancy strategies. Even the best intentioned reviews and regular update processes will miss important input from the very user community the taxonomy is designed to serve.
Because Oracle Universal Content Management schema also allows client side field validation, author-originators can be prompted to enter keyword data if this step is omitted. Several alternative methods for prompting contributors exist out of the box. Fields may be made globally “required” which means that all contributions will fail unless a non-null value is sent to the server for the field. A null value will produce a standard alert:

Alternately, rules may be set up and executed in the context of an Oracle Universal Content Management Profile that conditionally require field data. Furthermore, the Profiles capability allows much more granular evaluation of field data over and above simple null / non-null considerations. Additionally the message prompt to the contributor is fully extensible allowing for more meaningful messages. Even the addition of several more words in the contributor prompt can mean the difference between a user typing in nonsense characters in order to satisfy the non-null requirement of a globally required field and a user providing meaningful, democratic evaluation ready terms.

Finally, other author-originator keyword generation tactics may be implemented, including pre-set user or contribution folder defaults, automated keyword extraction and human validation with the Content Categorizer feature of Oracle Universal Content Management, or rule-based keyword derivation.

This, however, is only the very first step. Once keywords are entered, whether manually or automatically, they are stored in the database and exposed through the Schema API and SOA capabilities. With a simple component the keywords may be aggregated, tallied, and placed in a &lt;DIV&gt; tag to be displayed as a cloud
(as depicted above) in any web-enabled context and UI the enterprise desires. If
display is not the primary objective, then the keyword aggregates and tallies may be
queried through a Web Services or JAVA API call and leveraged for additional
relevancy enhancements and other API calls.

The keyword cloud depicted here was created with Oracle Universal Content
Management’s Idoc Script server side scripting language. It is emplaced simply by
calling a single Idoc Script include:

```html
<$include divKeywordsCloud$>
```

This include is evaluated by the server prior to delivery to the browser. The include
referenced contains two other includes and a service call.

```html
<@dynamichtml divKeywordsCloud@>
<$include typeCloudCSS$>
<$executeService("GET_CLOUD_FOR_DIV")$>
<$include KeyWordCloudContent$>
<@end@>
```

The typeCloudCSS inserts the style sheet declarations used by the cloud to govern
font color and size. While the keywords themselves are signifiers of the content in
the system that the user is able to “see” if she wants to, the cloud’s styling is the
primary means by which democratic signifiers are communicated to the user. To
indicate relative popularity size and weight are used. Bigger and heavier signify a
larger tally. To indicate the most popular keyword used, the style sheet offsets the
color, in this case, light blue. Finally, a styling hover effect is implemented to
indicate a drill-down capability has been incorporated into the cloud. Clicking on a
term immediately executes a search for items containing that particular keyword. In
this way, items sharing keywords are dynamically associated without any additional
author-contributor input. As a further enhancement to a basic keywords cloud, a
keyword constellation may be implemented where all keywords from items having
at least n keywords in common are highlighted. When clicked, all highlighted
keywords are passed as parameters to a search and the refined and highly relevant
results are returned to the user. A user whom, remember, had no necessarily
preconceived ideas about what she wanted.

The service call in the code sample above leverages the inherent SOA of Oracle’s
ECM applications. In this case, the script kicks off a JAVA API call which
executes the service:

```java
GET_CLOUD_FOR_DIV
```

This service has two jobs, to execute a query against the database and return the
results of the query in a JAVA result set object. The executing service and query
may be relatively simple, akin to a select and group query, to relatively complex,
implementing security rights in a where clause and merging distinct results with
results from count tallies into a single result. Regardless of how simple or complex
the query and any other methods executed in the service are, the results are ultimately made available to the display script that is part of the third include:

```
KeyWordCloudContent
```

The tallying of the keywords is where the power of the democratic strategies are realized. Sets of keywords falling below a pre-determined minimum threshold can be omitted from the final aggregate. Depending on needs, the remaining keyword tallies are subjected to some basic mathematical calculations to determine items difference, distribution and spread. Difference is the tally of the most popular keyword minus the tally of the least popular but still included keyword. Distribution is the difference divided by a pre-determined distribution denominator. This will ultimately determine how much variance there is in the tallies of individual keywords in the final aggregate. Larger distribution denominators mean less variance. Variance then, in visual presentation of keyword aggregates, governs the number of different styles that used (e.g. small, medium, large). Spread is the tally of unique variance sets (e.g. 3 small, 12 medium, 2 large) within an aggregate.

The capture, tallying, and presentation of author-originator keywords in Oracle Universal Content Management is not a complex endeavor. Yet the payback in efficiency, user experience, and user adoption is high when this strategy is combined with social tagging and the robust native full-text and metadata indexing that are part of Oracle Universal Content Management. Furthermore, this step is a necessary step in the journey to realizing both a highly relevant system and a dynamically semantic ECM system.

**Social Tagging in Oracle Universal Content Management**

The ways of implementing social tagging strategies in Oracle Universal Content Management are nearly limitless. The only requirement is to be able to capture the tags and the unique identifier of the content to which they apply. Additional enhancements that Oracle Universal Content Management facilitates are out of the box features of the schema API. This includes the following:

1. Automatic tag creation time stamping. Time stamping of tag creation enables tracking of the progression and evolution of tags and the nuance of meaning (ontology) and purpose (teleology) over time.
2. Automatic tag modification time stamping. Time stamping of tag modification enables micro tracking and capture of first impression, gut reactions, and overall usage.
3. Capture of response information. The capture and inclusion of any header and result-set data that is present is an inherent feature of Oracle Universal Content Management. Useful data such as user name, the page on which the tagging took place, the search query run by the user, the number of results returned, the complete result information etc. is available.
Furthermore, this header and result-set data may be leveraged to perform even more complex operations before or during the tagging event. This means that, since the user name is present, we could query for her most used tags for similar content items and return them to her while she is tagging as suggestions.

4. Extensibility. Because schema stores information in database tables, presents the data in views, and associates the data via a rules in a relationship engine that is system administrator defined (as opposed to DBA defined), and because the schema API knows how to dynamically access this information, any information may be captured and “mashed up” with the tags. This provides an exceptionally rich interface. To be sure, intentional forethought and clear goals are required when designing a tagging schema. However, this is not bad. Too many tagging strategies available today lack a purpose or goal and wind up being little more than yet more form fields that users are burdened with.

As with the above democratic keyword examples, the tagging example described below is intended only as a springboard for thought and development rather than a complete tagging solution.

For tagging to be social the tagging user interface must be accessible to the user where he consumes the content. Different contexts will place different demands on the implementation of the tagging solution. Public facing web sites have an anonymous user context where content is typically accessed by browsing or drilling down through a predefined navigation structure. Public facing sites run a much larger risk of ingesting irrelevant tags from passers-by. Intranets and extranets have a context (generally speaking) where users are known and content is searched for as much, if not more, than it is browsed or navigated to.

In the example, a tagging UI is emplaced on a search results page for each result. This allows content-specific micro clouds to help guide the consumer as they nuance the result. More helpful than metadata, such micro clouds give consumers a much better clue as to what it is to which the result pertains. However, including the actual tagging input in the same UI is for convenience and quick comparison only. Oracle Universal Content Management capabilities include the Dynamic Converter (DC) which, among other things, allows content and pieces of content to be chopped up, reassembled, mashed-up and re-presented to the consumer in whatever dynamic format or context is demanded or requested or triggered. An obvious example of a tagging DC template would be one that mashes up the content (or an abstract of it) with
this kind of a tagging UI. In that way, consumers could see how others have tagged the content while being enabled to add their own tags.

To provide a seamless user experience, Oracle Universal Content Management’s component architecture and SOA is again leveraged, this time, making a SOAP based service request and executing both the tagging and the retrieval of new tags via AJAX. This is important, especially in a results oriented environment such as in this example. It would be a poor web 2.0 tactic to allow users to tag items but require that they either navigate away from their working context or re-run the entire search in order to execute the tagging function.

In this example, a simple form is created for each result on the screen with the following:

```html
<a href="" onClick="showHide('%dDocName%')">Tag This</a><br>
<div id="%dDocName%" style="visibility:hidden">
<form name="tagMe" onSubmit="runSoapRequest(this); return false;">
Tags:<input name="tags" value="">
<input type=hidden name="tagDocName" value="%dDocName%">
<input type=hidden name="tagUser" value="%dUser%">
<input name="go" value="tag it" type="submit">
</form>
</div>
```

In keeping with the web 2.0 mindset of “show-it-if-you-need-it”, a simple visibility toggle wraps the “Tag This” label. Clicking it shows the tagging form enabling the consumer to supply her own tags. The input field may take as many tags as the user wants the provide presuming that they are delimited (commas are common). Along with the actual tags, the form also submits the user ID (<$dUser$>) and the content ID (<$dDocName$>) as hidden values. These are submitted to the tagging function (runSoapRequest(this)) which implements a callback function to actually submit the SOAP request, retrieve, parse and paint the results (i.e. the new cloud) after the submission has completed. All of these actions take place asynchronously – the end user does not see a page redraw, the server does not re-run the query. The tags are supplied and the new cloud appears.

Also incorporated into this example is a content rating capability. This combines democratic style voting with the social aspects of ubiquitous tagging and presentation of the (inherently social) tag cloud. In this
case, users are voting on whether or not the item is helpful. Such democratic tactics are not uncommon. Amazon.com™ uses similar democratic tallies in conjunction with their customer reviews. Content rating mechanisms may be sophisticated or simple. In our example, the “Helpful” checkbox only appears for users who are not the author or the content item. Those users can indicate the item was helpful as they tag. After some minimum number of non-author users have indicated that the content items was helpful to them, the cloud displays a small symbol, a signifier that this has been found by the user community to be, in general, helpful. In our example, the symbol is a small “thumbs up”. When a user mouses over the thumbs up symbol, a small line of highlighted text appears reading “This item has been rated as helpful by at least 5 people”.

More sophisticated strategies can incorporate the idea of power users or experts and add the concept of “weighting” to both tags and content ratings. Tags and ratings from expert users or “power users” would be weighted more heavily than tags and ratings from anonymous or non-expert users. The strategies of user ratings and weighting hail back to the notions discussed at the beginning of this paper, that credibility is the currency of the web. With signifiers of content credibility, we can infer that the likelihood of relevancy is high, or at least greater than the surrounding content items. Content ratings are indicators of content credibility. Incorporating weighted responses from expert or power users leverages the credibility those users gained in achieving membership in those super classes. Rating algorithms that weight ratings from expert and power users infer that such users are better or uniquely suited to discriminate relevancy among all the results in a set. We, the consumers infer that, due to their earned credibility, “the experts” are well situated to help us find that most relevant piece of information we are after amidst all the results with which we are presented.

Features and Benefits

The preceding sections are but the briefest descriptions of the work that has already and continues to be done at Oracle in the areas of social and democratic folksonomy strategies in Oracle Content Management. From these idea springboards, the following strategies may be integrated into your organization’s relevancy discipline and their benefits realized. Additional aggregation concepts and benefits of the folksonomic approach are presented below.
General Clouding

Clouds are a convenient and intuitive way to present information. It should be remembered though, that keywords, tags, popularity rankings, usage metrics, and rating scores are just more data contributing to the problem of info-glut if they are not made useful to humans or applications or both. Consequently, the benefits of folksonomic inputs, whether keywords, tags, rankings, usage or ratings are realized only insofar as they are exposed. Exposure may mean that the data is included in a search index. However, with the anthropocentric focus of folksonomy in general, human-centered display most often is part if not all of the exposure. As such, clouding is presented in this paper as one strategy with real benefits for end users.

Because clouds are beneficial for information presentation, there is no implicit reason why democratic clouding cannot be implemented for any metadata field. I mention democratic clouding as opposed to social clouding because metadata fields are typically not open to all consumers to modify. Consequently, the clouds, while still providing nuance over an aggregate of data are still tallies. The power of clouds is in the presentation of an aggregate of ideas. A single cloud may signify an aggregate of ideas over an aggregate of content as in the keywords cloud example presented above. Alternately, it may present an aggregate of ideas over a single piece of content.

Search Results & Micro Clouds

As depicted above in the example, these are small clouds containing tags for each individual item. Presented in a search result set, they provide flavored or nuanced meta-information over and above what is presented by default. This facilitates human inferences of relevancy. Because the search engine has already run and provided its set of results back to the user, the user is typically left with little recourse to determine which is the “right” content aside from opening and scanning each item or making less than educated inferences from metadata such as title. Providing small “micro” clouds related to unique content items facilitates a humanparing down effort.

Nano Clouds

Dialing down even further we can envision nano-clouds where the aggregated tags and keywords spring from sections of unique content items. If the task of enabling and prompting users to provide keyword and tag feedback for content items were not daunting enough, the presumptive level of effort in tagging discreet sections of content is near overwhelming. As a result, while users may be enabled to provide social feedback, nano-clouds may be automatically generated as keyword aggregates from full-text indices. One may envision leveraging capabilities already present in the Oracle Universal Content Management toolbox, such as the XML Converter (also included within Oracle Content Conversion Server) and the Content
Categorizer tool to assist with this task. An item may be checked in to Oracle Content Server and passed to Content Categorizer which automatically converts the item to an interim XML format and may then apply an XSLT transform to that item. The XSLT can chunk sections for consideration or preclude sections of the item from consideration. Alternately, XML Converter which can denote sections of the item and chunk them via an XSLT post conversion process. Each chunk can then be fed iteratively to Content Categorizer for rules based suggestion and precipitation of keywords and concepts which can be stored in a metadata or schema field. The net result is that the Content Categorizer engine can precipitate out keywords to be stored in Oracle Content Server schema tables and aggregated into section relevant nano-clouds. Such clouds then become very useful for application based processing and generation of ECM inference engines. The implications for the automated creation of compound, complex and conceptually related but heterogeneous content through this kind of a content mash-up is profound. Imagine an ECM system that can autonomously assemble a real-estate contract from a boilerplate document, a spreadsheet of addresses, another spreadsheet of property valuations, an agent directory and a report pulled from a banking system with loan rates and amounts and payment schedules and then route that content through workflow for final editorial reviews and then signatures by all transacting parties.

Cloud Constellations

Cloud Constellations are, quite simply, presentations of aggregates of clouds. These may be presented as clouds of clouds or in 3-D modeling and interactive arrays or oriented spatially in 2-D. The purposes are variable and different benefits are promoted through different presentation modes. Clouds of clouds facilitate a drill down approach to locating content. For example clicking on a tag at the “top” level (say, “bread”) yields a second cloud of more specific tags that either also include the first one selected (e.g. “yeast”, “bakery”, “slang”, “money”, “musician”), or come from content “bakery”, “small business”, “Linden Hills”, “Great Harvest”, “FY 06”). In the first example, clicking on “bread” yields a cloud with terms from three distinct content items, one pertaining to a bakery, one pertaining to colloquial terms for currency, and one pertaining to the 1970’s soft rockers. In the second example, clicking on “bread” yields a cloud with terms pertaining to the workings of a local small business bakery.

Other, non-hierarchical constellations may be useful as well. For example, a 2-D cloud of clouds, if not too cluttered (e.g. cloud of micro or nano clouds) may be very useful for detailed comparison of democratic and social attributes of several content items. Indeed, in the search results example
above, the search results page itself may be considered a 2-D cloud of clouds where the individual micro clouds are oriented linearly in a page that acts as a cloud.

Alternate Navigation Paradigm

As an alternative to browse or query to location navigation paradigms, the 3-D oriented tag or keyword aggregates provide some unique opportunities. Because tag and keyword aggregates can function as containers of containers, navigation schemas may be built around keyword ontologies and tagged teleologies. Consider a navigation bar organically generated from the web page context containing the most popular (democratically) keywords. Clicking on a keyword might drop down instead of a list of additional options, a tag cloud. This would visually represent the folksonomy of the content in that page pertaining to the selected keyword. In this way, usage patterns and what is relevant at the time drives site navigation (at least in part). Web sites, in this way, respond to trends before marketers even know that a trend is developing. Matched up with legacy business intelligence applications and usage analytics, end users of a site or application can navigate to the most relevant content within a selected context. To be sure, this cannot be the only kind of navigation on a site, but as part of a comprehensive relevancy discipline, the benefits are at least interesting and at most, profound.

Tag & Keyword Constellations

Tag constellations are fascinating ways of associating tags and/or keywords to each other within a container, commonly a cloud. Tags or keywords from disparate sources are visually associated when they meet a certain minimum criteria. The criteria may be that at least two tags or keywords from different content items match. The visual highlighting effectively denotes a sub-cloud within the container. A nice visual example of this is found at the “Google News Cloud” site here. The constellation creates an ad hoc content association that required no human input. Such associations are useful tools to act as content promotion and suggestion engines. They can help to expose users to content that they may not otherwise have decided to investigate. Furthermore, tag and keyword constellations help expediate and alleviate the outlier risks associated with social tagging. While over time, social tagging will address errantly or oddly selected tags, tag constellations will associate outlier tags with more common tags thereby lending the outliers some context and nuance that may otherwise be missing.
**Suggestion Engine**

Suggestions engines such as those found on sites like Amazon.com™ and others have been implemented after careful study indicating that they work. The principles that govern these kinds of retail “soft pitches” also have benefits to offer users of ECM systems. Typically suggestion engines will be one of two styles, either they will suggest based on what the user has done in the past or they will suggest based on what other users who have accessed that same content items have done in the past. Because of the flexibility and extensibility of the Oracle Universal Content Management schema API the best of both approaches can be used. For named users, tags that they have provided for content in the system can be aggregated to form a personalized baseline “interest set”. That baseline set represents the user’s interest within the context of the Oracle Universal Content Management system. Against that baseline set, then, “interest spikes” can be found by analyzing the baseline set for tag patterns that precipitate out over short time periods (e.g. one or two weeks). Additionally, by comparing the interest set of one user to that of other users who have tagged n numbers of the same content (socially) or have significantly overlapping “interest sets”, suggested item or suggested page links may be percolated up to the user. Add in content and page access information from the Content Tracker tool in Oracle Universal Content Management as a filter and the result is that extremely relevant content, of which the user was probably unaware, is promoted to the user.

This can be done for both baseline interest sets as well as for the smaller interest spikes. Because a relevancy discipline must implement multiple and diverse strategies to get the right content to the right people at the right time, any advantage that may be gained by proactively getting the content to the user before she asks for it (e.g. search) is substantial. Combined with auto-generated nano-clouds, content items containing only pieces and parts that may be relevant to the user may be suggested, opening the possibility of introducing the user to entire new areas of content previously beyond her ken.

**Role-based Relevancy**

Hyper-personalization is a hallmark of web 2.0. In order to get the most personally relevant information to the individual based on information known about that person (i.e. user metadata) term - term or word - word relevancy should not be the only factor. User role - term relevancy is also considered. In addition to the flexible and extensible content metadata API Oracle Universal Content Management implements, it also provides a similarly flexible and extensible user metadata API. User attributes such as role, role description, access levels, security clearance, manager, organizational structure, and interests are evaluated against keyword and tag aggregates. Automated role-based relevancy calculations are performed that may leverage thesaurus style comparisons of user metadata information against tag and keyword aggregates. Thesauri themselves may be social
tag artifacts and be built up democratically - they get better with time and use presuming good intentions and adoption of the paradigm.

Other Considerations

Capture, storage, evaluation, aggregation, and presentation of democratic and socially produced relevancy information is only the tip of the semantic iceberg. Once this information is captured it can be used for other user experiences and intelligent calculations.

If users are wedded to their hierarchies (e.g. within visual nested folder metaphors for information placement like MS Windows Explorer) the ability to create and present dynamic hyper-personalized hierarchies emerges by displaying the folder order based on conditionally calculated democratic user tallies. The most used (i.e. most popular) folders and folder chains would be displayed at the top of the folder hierarchy. Popularity would be determined by a democratic tally evaluation of keywords and tags and usage statistics captured for each user.

Thinking Semantically

The Semantic Web project has as its goal, the defining of inter- and intra-data relationships for the purpose of facilitating nascent computer reasoning through inferences. Though still evolving, most current approaches (where they even exist outside of the academy) are akin to developing and publishing a centrally defined taxonomy. They are centrally planned, brute force approaches that suffer both from the scope of the task (massive) as well as technology that has not quite caught up to the vision (both in the ability implement and maintain schemas across enterprises as well as to process massive amounts of data within timeframes end users find acceptable). However, as we start thinking according to semantic web principles, it is evident that the content item is no longer the fundamental building block of the “C” in ECM. Rather, discreet data becomes the fundamental building blocks. Ideas and concepts become the building blocks. Ideas and concepts are malleable rather than static and hyper-personal; while denoting the same thing (in practice if not in poststructural semiology) they may connote different things to different people. This is one of the reasons why the emerging mash-up paradigm of web2.0 is so inherently exciting to us as people. At a deep level, we get it. This is how we think and infer and reason. Concepts and ideas should be freed from the prisons of documents and web pages and the data should be made available. As Hamlet would say, “Aye, there’s the rub”. For, to make the data available to applications, it must reside in some structure or format that applications can understand. Welcome RDF.

Yet there is no inherent reason why RDF cannot be socialized or socializable. Each constituent of an RDF triple (subject, predicate, object) may be aggregated into a cloud. Conversely, automatic ingestion of social tags into an RDF schema
will both facilitate the creation of a truly organic and evolving semantic web (or at least semantic ECM system) as well as greatly increase the richness of inferences that can be made. Computers would have the RDF schema in order to make rich and (most importantly) non-static programmatic inferences about the content while human consumers would have cloud aggregates from which we can make inferences about the content.

WHY ORACLE?

Integrated Relevancy Discipline

A discipline is an ongoing activity, not a commodity. In order to be disciplined, one must have the necessary pre-requisites in place. Oracle has the only truly unified enterprise content management platform. Oracle Universal Content Management enables organizations to manage documents, web sites, digital assets, information rights, records, retention, document-centric collaboration, legacy formats, scanned images, on a single platform. When resting on Oracle's database and data storage capabilities, Oracle's superior infrastructure means that all the pre-requisites are in place for a successful relevancy discipline. Because Oracle's content management solutions are part of Oracle Fusion Middleware, organizations with heterogeneous IT systems need not worry. Oracle’s content management solutions are hot-pluggable. This means that the solutions are infrastructure agnostic and can work with multiple third-party technologies. Oracle Universal Content Management's robust Web Services, J2EE, FileStoreProvider, and Adapter APIs along with the truly Services Oriented Architecture means that integrations are a snap and many come right out of the box.

Organizations' production of information is outstripping their ability to effectively find let alone utilize it. While semantic web strategies will eventually help automate parts of the relevancy discipline, the sheer scope and size of the semantic web project means that progress noticeable by the lay-person will be some time in the coming. Organizations can ill afford to continue to gasp and struggle under the tidal wave of information while waiting for a relevancy discipline to simply happen.

The goal of a relevancy discipline is achievable within Oracle Universal Content Management. Oracle Universal Content Management already incorporates full text and metadata indexing, search/retrieve, metadata display, thumbnailing, preview, dynamic conversion, and automated categorization. The democratic and social folksonomy practices are the only pieces remaining to integrate. The infrastructure is present with the Oracle Universal Content Management schema API, dynamic conversion, the ability to automatically suggest/pull keywords from content, the ability to automatically track access and usage rates, and persist unique user information. Finally, Oracle Universal Content Management has a huge advantage over World Wide Web implementations of folksonomies which are only partial at best: Oracle Universal Content Management's folksonomy offering is a fully integrated, fully realized and unified approach that is able to be leveraged and purposed dynamically across the ECM solution.
CONCLUSION

The information explosion has had dire consequences for the goals of relevancy. Getting the right information to the right people or systems at the right time continues to be a challenge. What the emergent web 2.0 mindset has generated are a number of tactics that share an inherently social aspect. Democratic practices foster user involvement for the purposes of tallying something or voting on something. Social practices foster user involvement for the purpose of enhancing content ontology and teleology. Together democratic and social practices for adding information to or about content is known as folksonomy. Oracle offers a folksonomic implementation strategy within Oracle Universal Content Management which leverages the power and capability of Oracle technology to provide folksonomic strategies that combine into an overarching relevancy discipline. This discipline has a number of real benefits across not only relevancy goals, but also business intelligence, analytics, user experience, and semantic web goals. Because of Oracle’s unified ECM platform and the folksonomic strategies envisioned and demonstrated here, Oracle Universal Content Management is the only choice for a fully realized relevancy discipline that works today.