Enterprise Knowledge Management and Emerging Technologies

Jonathan Grudin
Microsoft Research
Redmond, WA 98052
jgrudin@microsoft.com

Abstract

Improving management of information and knowledge in organizations has long been a major objective, but efforts to address it often foundered. Knowledge typically resides in structured documents, informal discussions that may or may not persist online, and in tacit form. Terminology differences and dispersed contextual information hinder efforts to use formal representations. Features of dynamic emerging technologies—unstructured tagging, web-logs, and search—show strong promise in overcoming past obstacles. They exploit digital representations of less formal language and could greatly increase the value of such representations.

1. Introduction

Knowledge management includes acquiring or creating knowledge, transforming it into a reusable form, retaining it, and finding and reusing it. Small organizations focus on knowledge acquisition; with few people and limited dispersal of knowledge, they face relatively few obstacles sharing or reusing knowledge. Large organizations, in contrast, have difficulty finding and reusing knowledge. Even determining whether the knowledge exists within the organization can be difficult. For example, a pharmaceutical company found that although clinical tests of a compound are expensive, searching for possible past test results of a compound would be more expensive than retesting some of them.

Digital technology has long seemed to be an obvious way to improve enterprise knowledge management. Information that is represented digitally and placed on an intranet can be accessed by anyone in the organization any time in the future. Document management systems, directories of personnel identifying areas of expertise, and other repositories are constructed and used in some circumstances. But their use has proven to be far more limited than many expected. They are expensive to create and maintain, limited in scope, and cumbersome to use.

Wanda Orlikowski's widely-cited study of a large-scale deployment of Lotus Notes nicely presented a vision of consultants sharing knowledge to avoid the need to constantly recreate it [15]. Despite the considerable sum invested in a technical solution, it failed. Individual consultants did not perceive enough benefit to undertake the considerable training needed to use the system.

This was by no means a unique experience. For over twenty years the authors have intermittently worked on efforts to develop design rationale and expertise location systems intended to fill knowledge management gaps. These efforts did not succeed as intended. The literature is strewn with other efforts that had limited impact [e.g., 14].

Assessments of these experiences often focus on a conflict between the tacit nature of much knowledge and the requirement imposed by digital systems that information be represented explicitly. This was a valuable insight, but the conflict no longer seems as potent as it once was.

Tacit knowledge is often transmitted through a combination of demonstration, illustration, annotation, and discussion. Little of this was represented digitally; representation that did occur was ephemeral.

This is changing with breathtaking speed. With a terabyte of memory is available for under US$1000, virtually all conversation and other forms of activity useful in spreading tacit knowledge can be captured and made persistent. This is not enough in itself—such information must be practically accessible and comprehensible. But the situation is changing, old assumptions do not hold, and in fact new opportunities are emerging. In this paper we present an argument and evidence that features of unstructured tagging, weblogs, and search show remarkable promise in overcoming obstacles to effective knowledge management in many enterprise settings.

Knowledge management research and development projects have pursued many approaches to capturing and reusing knowledge. These include creating document repositories; recording meetings,

conversations, and email exchanges; organizing discussions in document databases; and providing annotation systems. Key obstacles to success are:

1) Digital objects are difficult to find.
2) When found, objects are difficult to assess.
3) Systems are not strong at identifying people who can help find or assess objects.

The first two challenges—the difficulty of making effective use of keywords and other metadata, whether organized in a hierarchical ontology or simply added to documents, and the separation of objects from the context that defines them—are familiar, because they also limit the utility of paper repositories. They are receiving more attention in the digital domain due to the growing proportion and sheer volume of information that is represented online.

The next sections discuss these challenges and propose that the enterprise knowledge management logjam may be broken up by rapidly emerging and evolving technologies that are today mainly deployed publicly on the web: unstructured or collaborative tagging, as in applications such as flickr and del.icio.us, and weblogs, which are already seeing some internal enterprise use. The third hurdle may be the most difficult for technology to handle, but as outlined in the fourth section, effective use of search capabilities together with features of tagging and weblogs may reduce the need for human intermediation and make access to human expertise more effective and less intrusive when it is required.

Key characteristics of these technologies are that they 1) can be extremely lightweight, 2) make information and activity highly visible, 3) provide individual and group benefits, and 4) are grassroots, self-organizing phenomena. In enterprise settings the first three are necessary; it is too early to tell what degree of mixed bottom-up and top-down influences will be most successful.

2. Digital documents are difficult to find

2.1. Ontology and its discontents

At the heart of most knowledge or information management systems of any size is a set or system of keywords or classification terms or phrases. In the absence of highly reliable natural language analysis and recognition systems, descriptors are typically used. Examples are the Dewey Decimal and ISDN systems used by libraries and publishers, the descriptor and keyword system increasingly mandated by ACM for conference papers and journal articles, and the Dublin Core structured tagging system for document repositories [9]. Categorization systems can work, but they require considerable effort to establish, maintain, and use.

Creation of metadata requires that the creators of objects, or people working on their behalf, put in the effort to add metadata for the potential benefit of others who generally remain unseen and may in fact never materialize. Object creators often have little incentive to generate metadata. It has often been a cumbersome process of accessing and filling out a form for each object a problem made more difficult by the need to generate labels that will be consistent with past labeling and useful to others. This problem can be partially addressed by devising an overarching classification system. Such systems require considerable effort to create and maintain. For example, the ACM classification system is so sparse in rapidly-changing areas that authors are forced to rely on author-generated keywords anyway. Internal enterprise taxonomy creation efforts with which we have been involved have not succeeded.

A second set of difficulties is encountered by the users of metadata. If there is a formal system, they must find it and obtain an overall sense of it. Identifying the best places to look can be difficult, and as noted, the centrally maintained taxonomy may not be useful. Items must be looked up in indices, such as library catalogs, or memorized, as in Latin names for species.

Research consistently shows that when author-generated metadata is relied upon, people differ in their use of terminology to a startling degree (see for example "The vocabulary problem in human-computer interaction." [5] There is no agreement on the most appropriate terms. I may label a set of London-area vacation photos "England," you may label a similar set "Great Britain," and she may label them "UK." I may even forget which label I used and end up searching files under "London," not finding my pictures, much less yours or hers.

An alternative, a pre-defined hierarchical set of labels, has weaknesses analyzed by Clay Shirky in a brilliant essay, "Ontology is Overrated" [16]. Such top-down approaches must, in his terms, read users' minds and predict the future, among other challenges.

Given the mixed outcomes with widely-deployed systems, it is not surprising that metadata has not proven wildly successful in improving enterprise knowledge management. Hope springs eternal: An organization that has a relatively small set of concepts to handle may envision a high collective benefit to adopting an ontology, but such systems die of neglect without substantial, expensive administrative efforts.

2.2. Unstructured tagging: A self-organizing solution?

Unstructured tags are words or short phrases associated with objects for the purpose of identification or retrieval. They are essentially keywords, but etymologically, keywords suggest a link to text ("key word"), whereas these tags are often applied to objects such as photos or URLs where the tag itself does not appear at all.

Unstructured tags are the driving force behind immensely popular sites such as flickr, an archive of around 50 million photographs, and del.icio.us, an archive of URLs. Both allow anyone to place their personal information on a freely-hosted web site.

They are very lightweight, capitalizing on what has been learned about building simpler interfaces for handling personal digital photos and URLs. In fact, *a key reason for their success is that people use them for personal information management.* Being web-based, they have a significant advantage over objects stored on a hard drive, in that they can be accessed from any machine on the Internet. I can bring up family photos or URLs of interest on the computer of someone I am visiting. They also serve semi-private purposes, such as allowing me to share photos with distant family members much more easily than burning them onto a disk.

Many people restrict tags to labels with personal meaning, such as "August 2005 Vacation" photos or "To Read" URLs. But with no limit to the number of tags one can add, I can also make them available to others, and perhaps be rewarded by comments and recognition by others. And this is done. The blend of personal and collective benefit is captured in the definition of del.icio.us as a "social bookmark system" and the somewhat controversial tag 'folksonomies' given to these sites. Figure 1 shows some of the flickr photos tagged "londoneye."

In *The structure of collaborative tagging systems* [6], Scott Golder and Bernardo Huberman discuss the costs and benefits of author-generated metadata. They stress problems, including polysemy (one word with different though related meanings), homonymy (one word with unrelated meanings), and synonymy (the 'Tower of Babel' problem noted above, people applying different terms to the same object. They also show data to support the fact that the evolution of meanings over time affects there 'self-organizing' systems in ways similar to those shown by Shirky to plague top-down hierarchies.

The features that correct for these problems are the extreme ease of seeing what terms other people are using, and the ability to use multiple tags. This is brought out in Figure 1 by the "related" and "see also" lists, and in both flickr and del.icio.us by lists of currently popular tags (Figure 2).

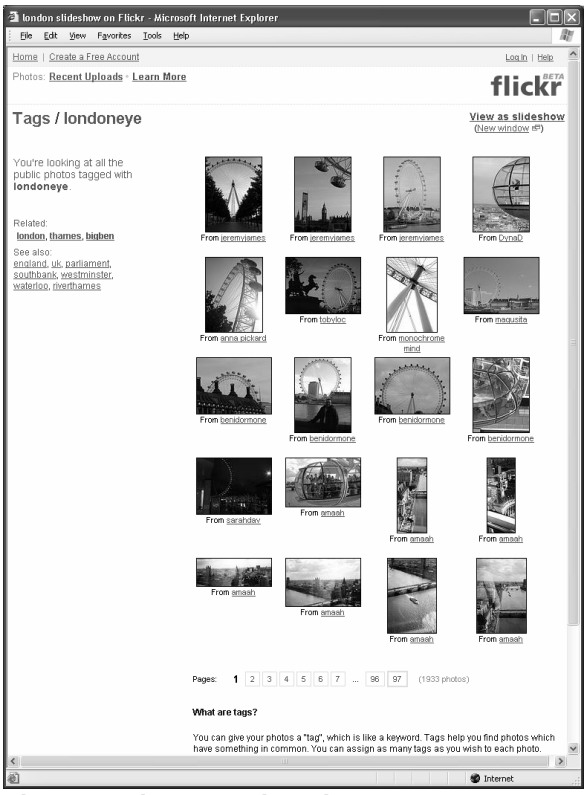


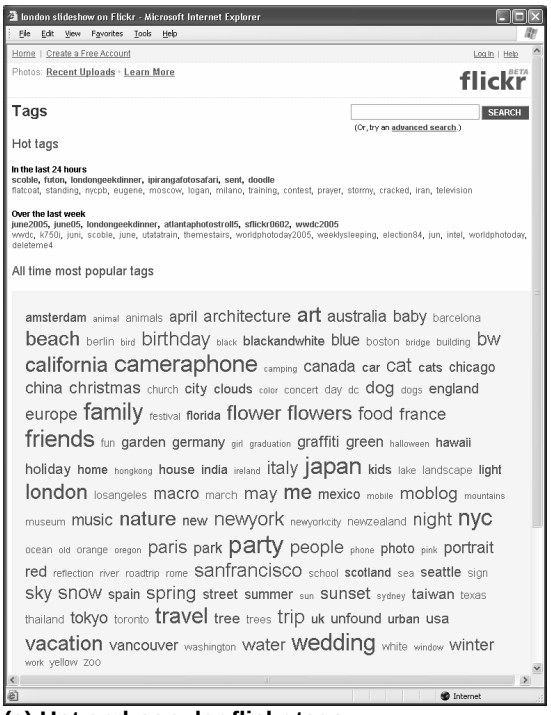
Figure 1. Photo tagging site.

One can also quickly contrast existing use of different tags to help choose tags for new photos. Table 1 shows the number of photos with particular tags on flickr on 8 June 2005. Depending on your goals, you can use multiple tags, stick to popular tags, or select those that will help with personal retrieval.

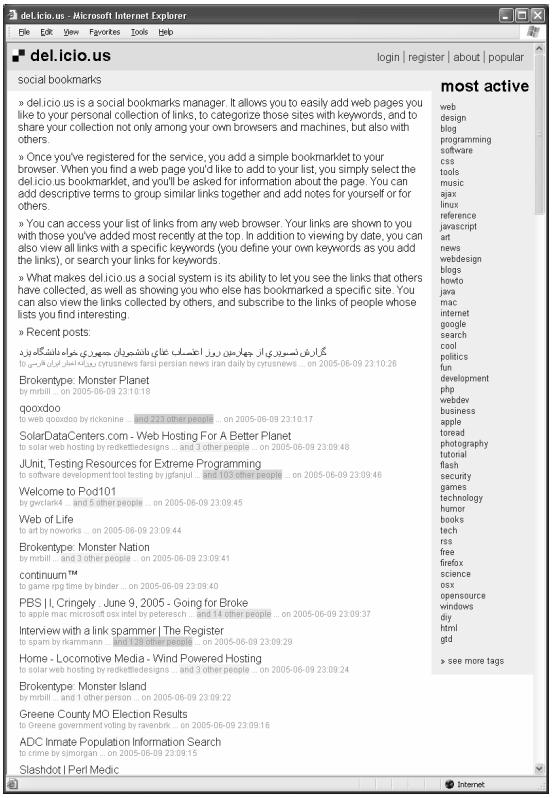
Table 1. Frequency of flickr tag use.

flickr tag	Number of photos
England	39,658
UK	33,475
UnitedKingdom	4,709
Britain	2,250
GreatBritain	1,394
London	91,701

Note that unlike structured tagging systems, unstructured tags are not inherited automatically. London is part of England, but photos tagged London outnumber those tagged England. Inheritance cannot be automated due to polysemy—some could be photos of Jack London or London, Ontario.



(a) Hot and popular flickr tags.



(b). Most active tag list on del.icio.us. Multiply cited URLs are highlighted among recent posts.

Figure 2. Making activity visible.

The bottom-up structure thus leads to some messiness, but also has great power. Users can very rapidly scan items in the case of flickr and look at cumulative endorsements of URLs in del.icio.us. They can also look at subsets of items grouped by creator. The ease with which a viewer can leave a comment or question is another significant enhancement.

A similar repository of author-tagged items visible across an intranet is a plausible way to make enterprise metadata usable and useful. To what degree an entirely self-organizing system of this sort will work is an unresolved issue, returned to below.

3. Making sense of found objects

3.1. Document repositories: content without context

Heavyweight document management systems have been available for some time. The Web has made document repositories more accessible and easier to use. The pioneering GMD BSCW system has hosted documents since 1995 without charge for users anywhere and licensed the software for use behind firewalls [1]. Microsoft's Sharepoint and IBM QuickPlace are recent applications that provide web portals to document repositories.

Over time, such repositories have become more versatile, as well as quite easy to create and use. Any project can easily archive its documents, facilitating immediate collaboration and later reuse. So, why have such repositories not solved the knowledge management problem?

Consider the challenges faced by people seeking information that is contained in such a repository. They must find the relevant repository, obtain access to its contents, and then interpret the collection of documents stored there. Each step can be difficult, but the final step can be the most challenging. A document repository that contains most or all of the structured content produced by project members is typically undecipherable to people outside the project, to people who join the project after it is underway, and even to the original project members after time has passed.

A folder of software programs prepared for use by other people is usually accompanied by a 'ReadMe' file that explains the purpose of each of the other files. Document repositories generally lack a resource equivalent to a ReadMe file, and for obvious reasons. Creating and maintaining such a file would require extraordinary discipline on the part of the team. But without it, anyone new to a project who examines its

document repository lacks the context needed to understand each document. Which documents were central, which were background? Is this the latest version? Was it a complete draft or was it made available because one section was finished? Are any unresolved issues associated with it? Has it been superseded by another document? A person sifting through such a repository often tries to track down project members who can remember enough to help.

Creating such a ReadMe file can be viewed as analogous to commenting code. Programmers add comments to their code to provide essential context. They hope that these comments will help other readers of the code understand its purpose, structure, and approach. Programmers may become more willing to comment code when they realize that they themselves have difficulty understanding their own code months later. Understanding the content of a large document repository is similarly challenging.

Of course, project members do write comments about their documents, but the comments are not captured in project repositories in a reusable form. An author typically sends email to other project members when adding a document or new version to the repository. This email serves an essential purpose at that moment, but is not helpful to people who join the team later, people who are on other teams that develop a need to know, or even to team members who have difficulty finding the relevant email months later. The knowledge is not well managed.

3.2. Project weblogs: providing context

A project blog [19, 19] linked to a project document repository is a potential partial solution to this problem. Instead of sending email when creating or revising documents, project members simply add an entry to the blog. This requires *less* work than sending email and has other attractive features discussed below. The result is an easily-skimmed chronologically-ordered record of the important information events of a project. These include status changes of project documents in the associated repository. Project blogs are collaborative; any team member can add an entry.

Project members can choose to be automatically notified by email when a new entry is added to the blog, and through this means they learn when documents are created or modified. The project weblog communicates in real time among project members and also serves as the project's minutes; it reminds, educates new team members, and is a resource for people on related projects.

Project weblogs are in limited use today. They have not received the attention given to other types of blogs. Of course, outside of high-tech and media companies, blogs of any kind are still rarities within government or private industry, where most projects are found.

That said, corporate blogs are getting attention if one looks for it. In a remarkable comment on the pace of technology adoption in today's world, Gartner Group places corporate blogs (and also wikis and RSS) well along a technology "hype cycle" even before most people have heard of them. Gartner considers them past the "peak of inflated expectations," approaching the trough of disillusionment, with maturation into productive use expected within a few years [4].

A thoughtful and well-researched popular press treatment is Edward Cone's "Rise of the Blog" in CIO Insight [2]. It discusses internal project blogs and other corporate uses of weblogs.

Next we position project weblogs in the larger blogosphere, emphasizing enterprise uses of the form and focusing on primary uses. For a more detailed analysis of corporate uses of weblogs, see [11]. Then we identify key features of project blogs and briefly describe their use in three projects.

3.2.1. Weblog categories and uses.

Diary-like blogs, authored by one young person primarily for reading by a small set of friends, are by far the most prevalent type of blog. As a result, millions of people will bring multimedia authoring skills into workplaces, including the skill of engaging readers through personal revelation.

A-list blogs, written by journalists and others, command the most media attention. They can be a good source of information on events, products, and trends.

Watchlists, reports of any appearance of a text string in any blog, are a powerful way to see how a product, organization, name, or topic is being discussed around the world.

Externally visible employee blogs typically discuss both personal and work life. The legal and public relations risks for a corporation can be offset by the benefits of putting a human face on an organization or product. A growing focus of media attention, these genre-blending [8] employee blogs illustrate how experience with diary blogs can be put to use in workplace settings.

Project blogs are the internal equivalent of externally visible employee blogs. Authored by multiple team members, they tend to focus exclusively on work.

3.2.2. Technical characteristics of project blogs.

Project blogs are very lightweight. In five minutes, someone with no programming skills can set up a freely hosted web-interface blog (e.g., blogger.com) and a password-protected document repository (e.g., BSCW) with reasonable interfaces and functionality. This includes the time to send invitations to other project members to join both. Posting an entry in the blog can be quicker than sending email to a distribution list, because one need not type an address.

Project blog entries are easily accessed. Rapid publishing, syndication, and aggregation features handle real-time notification, making a project blog as convenient to read as email. There is no need to access a web page to pull in blog entries.

Project blogs are chronologically sequenced. People are adept at reasoning unconsciously about chronological information. We know that information of different kinds ages at different rates. We know how to interpret waxing and waning of activity over time. For example, a sequence of closely spaced entries by several authors reveals the intense collaboration that contributed to a document. Repositories often provide some chronological information, such as when a document was last updated, but finer-grained chronological information is not readily visible.

Project blogs are easily skimmed. The chronology and the single-page format enable rapid skimming and reviewing. Consider a project that uses email to send information about new or revised documents. Retrieving and scanning the content of these email messages is not easy. In an ideal situation, all relevant email messages, including those sent by the person searching for information, were saved in a single folder with no other irrelevant messages. Even in this ideal circumstance, each message must individually be opened to access its contents.

3.2.3. Behavioral characteristics of project blogs. Two important features of project blogs cannot be enforced by the technology. They are social conventions to be learned and adhered to by project members for the technology to be most useful, just as excessive posting or abusive flaming must be avoided to fully benefit from email use in organizations.

Project blog entries are concise. To facilitate rapid skimming and reviewing, lengthy drafts or detailed information should be placed in the document repository with a mention and link in the blog, and items of highly ephemeral interest should go to a group email distribution list.

Project blogs, like most other weblogs, are single-voiced. Most diary and A-list blogs have a single author and thus a single voice. Several aides who author a politician's blog, or like-minded pundits forming a group blog, also focus on what they have in common. Most externally-facing employee weblogs are single-authored, although one can find "product blogs" that have multiple authors, often with individual bylines, or that involve uncredited authoring, reviewing, and light editing.

Retaining document drafts and capturing the communication around debates and disagreements are often important, but arguably the full text does not belong in a project blog. For one thing, it conflicts with the appeal to be concise: Skimming is more difficult when a blog contains drafts, debate, and consideration of options. Disagreements or options not followed that seem noteworthy at the time are rarely so interesting after they have been resolved. A discussion that might be of later value can be collected and stored in the associated document repository with a summary and link contributed as a blog entry.

Commenting and responding to comments is a less obtrusive way to have discussions within a blog [3]; limited commenting in a project blog could be useful. Because it will slow review, or be missed in review, one might anticipate that it be used sparingly, but this is an open question.

Because blog technology itself does not discourage debate or encourage concise, single-voice writing, it will be interesting to see whether project teams discover its value and employ social pressure to encourage it. We suspect that it will be crucial to the effectiveness of project blogs.

Internal project weblogs are widespread at a number of high-tech companies. The next section focuses on one that we observed closely and two with which we have had direct experience. Although short of formal study, these experiences indicate that project blogs can be effortlessly useful, and revealed some of the points noted above through trial and error.

3.2.4. Experiences with project blogs. Two anthropologists working for the Windows development team at Microsoft used a weblog to communicate their experiences in the field to members of their team and other interested parties within the company [12, also discussed in 7]. Lacking a production weblog-hosting tool, they reproduced many blog capabilities (Figure 3). Clicking on the image at the upper left launched an audio-and-digital-photo account of the site they visited the day before. In the upper right is information about one of the people who participated. Chronological

entries allow readers to review earlier posts. Links at the lower right are to documents. This engaging, easily accessed multimedia format was a successful communication medium.

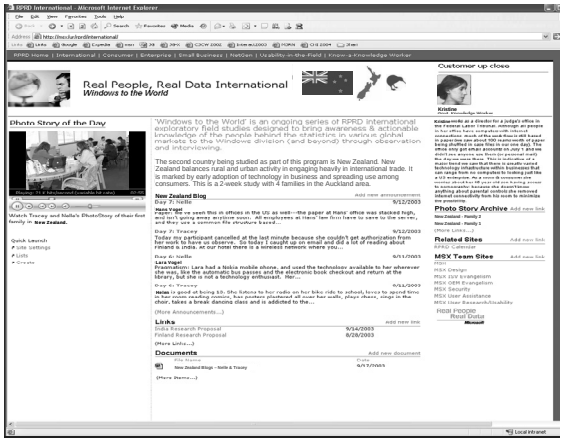


Figure 3. Project blog for an ethnographic study. (Details changed to insure privacy.)

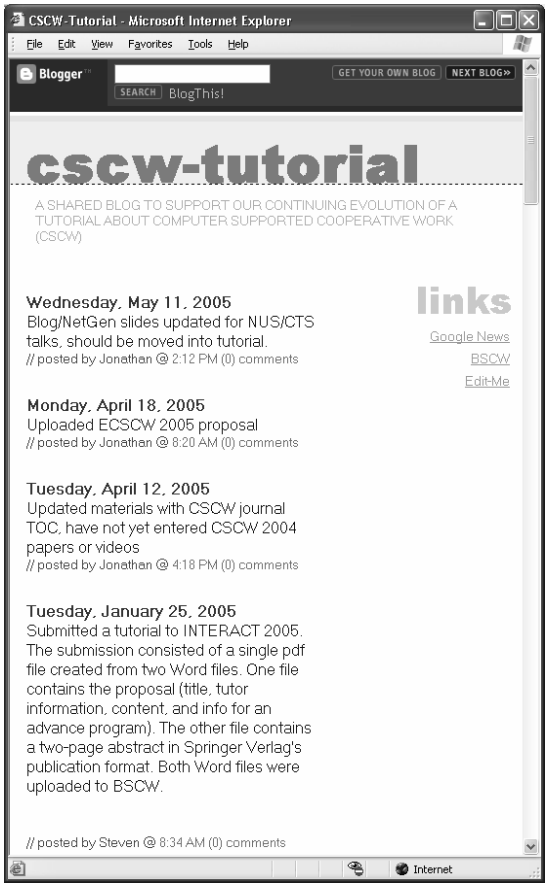


Figure 4. Project blog. "BSCW" link on right is to a repository for associated documents.

In the past year we created two simple blogs, one for an ongoing two-person project and one for a new three-person project. These real, albeit low-traffic, projects provided a low-cost opportunity to try the approach. In each case, one participant was initially skeptical of the utility of the approach, but became convinced.

The first is an ongoing project of several years duration that involved the creation and maintenance of a set of interrelated documents for a full-day tutorial given about twice a year at conferences. Prior to using the blog and document repository, each participant kept a copy of each jointly authored document. Most project communication was through email, which we did not manage well, saving it interspersed with unrelated email. Because of forgotten email updates or misplaced document versions, we frequently had synchronization problems leading to parallel documents, wasted time, and some embarrassment. On a few occasions it might have been useful for us to let a third instructor or a prospective student review some of the project history, but that had not been feasible.

We established a BSCW document repository and a project blog (Figure 4). Whenever one of us modified a document, we left the new version in the repository and posted an entry explaining what had been changed in the blog. Sending status updates to the blog was easier than sending them by email, an incentive beyond anticipated efficiency and reuse. We also posted action items, suggestions for future consideration, and snippets of information that we expected to be useful later. Discussion about the project that was deemed to be of ephemeral interest occurred off-blog, in email or meetings. On occasion one of us summarized such discussions in a post to the blog.

Activity in this long-term project is cyclical, with long periods of inactivity followed by a few weeks of intense work. Re-establishing context had been a challenge when returning to the project. The blog was effective in resolving this problem. In a few minutes of skimming we can readily identify our most recent versions of each document and are reminded of our thoughts about future directions posted months earlier. The infrastructure also enables us to engage other people in the project. Coordinating via email was barely manageable in a two-person project and would be more difficult for a larger group. We chose to make the blog public, but our BSCW site is password-protected.

The second project blog was used similarly, but the experience evolved differently. One team member initially saw the weblog as a replacement for all email and posted lengthy proposals and draft documents directly in the blog. This rendered it cumbersome to the point of being useless—this soon-outdated content was in the way when we skimmed for less ephemeral

information. Before long we deleted the superfluous material, moved discussions to email, moved draft documents and some email threads to the document repository, and the system seemed to work smoothly.

3.3. Project wikis

A versatile alternative to a project blog is a project wiki. Wikis provide more structure, are not wedded to the reverse chronological posting sequence, and are open to authorship by all team members. Wikis appeal to managers, who deal with information structured as documents, slides, and spreadsheets. But wikis lack some of the advantages of blogs in a project context. They are not as lightweight—they require up-front design, may require restructuring, and generally demand some ongoing oversight. Information that doesn't fit well into the overarching conceptualization may be omitted. Someone not familiar with the layout logic can have difficulty browsing; determining key features such as the authorship or currency of material can require digging. Finally, distributed authorship can reduce the incentive to contribute, creating a Prisoner's Dilemma. A project wiki is particularly appropriate when a deadline drives participation and a clear division of labor is in place; for example, it could be a great choice for planning a conference.

Knowledge management requires merging structured and less structured, more conversational information. Many proposed solutions, including wikis, stress efforts to add structure and filter out conversation (although wikis often do provide places for discussion). For projects with sufficiently dedicated management, wikis may work, as can other labor-intensive knowledge management approaches.

4. All you need is search

"Nothing you can know that isn't known, nothing you can see that isn't shown…" – Lennon/McCartney

As time goes on, there are fewer questions you can ask that haven't been asked before, online, *in the same way you would ask them*, with questions and answers captured for the ages by tireless web crawlers.

Rapid progress in improving search engines enables them to outmaneuver classification systems based on keyword matching. Stylos and Myers [17] found that people use search engines to find answers to questions more efficiently than they can by directly searching the pertinent online documentation that contains the answers. They note "One of the reasons that Google was effective at this task was because it worked well even with the use of non-expert terminology." If the documentation calls it "variable alpha compositing," but you don't think in those terms, a search for "fading image" may get you what you want.

This is one way that search engines that home in on persistent conversations (as well as structured information) address knowledge management challenges. In the introduction, we listed three challenges: finding a relevant repository, interpreting the information in it, and obtaining expert assistance. Search can assist with all of these, but it is currently hampered by a short-lived phenomenon: Powerful tools are available for searching the web and for searching an individual hard drive, but generally not for enterprise data, for searching within a firewall. This is one reason people use web searches to find answers that reside on an intranet.

Effective intranet searching is hampered by the system heterogeneity that exists within most enterprises that are not small and young. The web presents a relatively uniform terrain, as do hard drives.

Attention is turning toward bridging this gap in search capabilities, and we expect that the results will prove extremely important for enterprise knowledge management.

Software companies such as Verity are working to bridge this gap in search capabilities, to span the variety of formats of digital information across an organization. Researchers are looking closely at how people retrieve information for personal reuse [9]. In 2005, for the first time in fourteen annual meetings, the NIST-sponsored Text Retrieval Conference (TREC) included an Enterprise Track, "the purpose (of which) is to study enterprise search: satisfying a user who is searching the data of an organization to complete some task." [18]

It behooves us to follow the examples of the researchers cited above, and look at how people use web searches, *to extrapolate how organizational behavior will be affected when enterprise search appears*. Examples: With a computer affected by spyware or adware, a person identifies unknown executable files, searches for discussions about them on the web, and finds whether or not they are harmless and for those that are not, how to remove them; seeking a magazine article published some time ago, a person searches for the author and topic and finds an online discussion from the time of publication with a still active link to the article.

When enterprise search becomes as easy as other searches, structured data will be more useful, but equally to the point, the value of persistent online

conversation will be greatly enhanced through such opportunities for reuse.

Image recognition lags text recognition, but it is also an area of active research and development. By using text associated with images and accepting false positives that a human can sort through, applications such as Google's image search are useful. This too will become available within enterprises.

To summarize, as intranet search becomes as powerful as web search is today, skills honed in using the former will apply, and the value of persistent conversation within the enterprise will increase substantially due to being discoverable. This will aid in providing context for (and therefore value to) structured information, which will also be more easily accessed. And there will be a secondary effect on expertise location. It will be easier to identify relevant experts in an organization. It will be less necessary to bother them, because more answers will be found directly. And it thus may be more acceptable to contact those experts, because the need will be legitimate, the person will be appropriate, and the answer when obtained can join the online repository and thus be less likely to be asked repeatedly, a significant disincentive for experts to make themselves available for questioning.

5. Discussion

We have outlined how technologies that are coming rapidly into wider use can address obstacles to effective knowledge management. Features shared by these technologies are that they are lightweight, provide individual as well as social benefits, and promote visibility that includes views of social activity around information. As a consequence of these characteristics, they are on the whole marked by a bottom-up, self-organizing growth path.

That said, enterprise use will not come without planning. For example, free blog servers and unstructured tagging sites exist, but enterprises will not entrust all information to external systems. Where organizations support internal blog servers they are used; where organizations do not, internal blogs may remain rare. For such reasons, we anticipate that features of emerging technologies will be incorporated into products such as document management systems.

These technologies, to varying degrees and in one way or another, represent moves toward exploiting informal practices, behaviors, and conversations. It has long been noted that digital technologies are more readily applied to manage explicit formal representations, whereas people rely on a mix of formal and informal modes when retrieving information. Approaches to applying digital technology to knowledge management have played to the computer's strength, focusing on formality--hierarchies, indices, uniform keywords and metadata categories, structured documents, and so on. This succeeds when the benefits are perceived by all to be great enough. Often, however, the perceived benefit is not enough to motivate the effort necessary to overcome the unnatural stress on formality. This is particularly true for individual contributors in organizations, who rely more on informal communication and less on structured information than do managers.

Unstructured tagging allows people to use their own language and see language generated by others. Project weblogs associated with document repositories can place structured information in a context that may have been previously represented informally and not retrievable with the structured information. Search can enable people to use their own language to find information often described differently by others, with one effect being the discovery of conversational translations.

By exploiting informal activity or language that has been digitally preserved, these tools increase the value of conducting such activity or discourse in a manner that is preserved digitally. The significance of this positive feedback loop is truly worth reflecting upon.

It is early to say how these technologies will evolve. Top-down structure can serve a purpose. Self-organizing structure, derived by the system from author-generated information, may not be efficient, if it requires too much trial-and-error learning and repair by each user. It is unclear how the tradeoffs between bottom-up self-organization and top-down management could play out in different contexts.

When are blogs the right tool, when are wikis? Wikipedia has grassroots activity at the core, yet some structuring and management are also evident. When does it become useful to discern clear patterns in activity and strengthen them or provide overarching structures? When is it more effective to place more weight on the formal side? Each of the natural sciences progressed from a descriptive phase to efforts to categorize phenomena and then to a formal theoretical stage. Although it was crucial not to push ahead prematurely (as did astrologers, alchemists, and exorcists, *inter alia*), formality proved to be invaluable.

The answers to such questions will no doubt vary with the context and over time. The tools we build should anticipate and abet the need to restructure organically self-organizing systems when their growth becomes constrained. When we drafted this paper, the

limits seemed visible. In June, for example, there were 70,000 photos tagged 'london' on flickr. By September there were over 200,000, and it became reasonable to question whether a limit would be reached. Was a tragedy of the commons approaching, too many objects for there to be an incentive to provide more, and no way to reorganize the information entered? But then flickr introduced clusters, subsets organized around items sharing three tags, providing a refined level of self-organization. They introduced pools, allowing another avenue for users to organize. These are still early days.

6. Acknowledgment

Steven Poltrock, Lilia Efimova, Danyel Fisher, David Fono, Robin Jeffries, and Kate Raynes-Goldie and reviewers contributed to this work.

7. References

[1] Bentley, R., Horstmann, T., Sikkel, K. and Trevor, J., "Supporting Collaborative Information Sharing with the World Wide Web: The BSCW Shared Workspace System," *Proceedings of the 4th International WWW Conference,* 1995, 63-74.

[2] Cone, E., "Rise of the Blog," *CIO Insight,* 5 April 2005.

[3] Efimova, L., "Beyond Personal Web Publishing: An Exploratory Study of Conversational Blogging Practices," *Proc. HICSS'05, 2005.*

[4] Fenn, J., Linden, A., and Cearley, D., "Emerging Technologies Hype Cycle 2005," Gartner Group, 2005. http://www.gartner.com/teleconferences/attributes/attr_129930_115.pdf.

[5] Furnas, G. W., Landauer, T. K., Gomez, L. M. and Dumais, S. T., "The Vocabulary Problem in Human-Computer Interaction," *Communications of the ACM,* 30, 1987, 964-971.

[6] Golder, S.A. and Huberman, B.A., "The Structure of Collaborative Tagging Systems," Technical report, Information Dynamics Lab, HP Labs, 2005. http://www.hpl.hp.cpm/research/idl/papers/tags/

[7] Grudin, J., "Communication and Collaboration Support in an Age of Information Scarcity," in K. Okada, T. Hoshi, and T. Inoue (Eds.), *Communication and Collaboration Support Systems.* Ohmsha, 2005.

[8] Herring, S.C., Scheidt, L.A., Bonus, S., and Wright, E., "Bridging the Gap: A Genre Analysis of Weblogs," *Proc. HICSS'04, 2004.*

[9] http://dublincore.org/

[10] Jones, W. P., Bruce, H. and Dumais, S. T., "Keeping Found Things Found on the Web," *Proc. CIKM 2001,* 2001, 119-126.

[11] Jüch, C. and Stobbe, E., "Blogs: The New Magic Formula for Corporate Communications?" *Deutsche Bank Research, 53,* 22 August 2005.

[12] Lovejoy, T. & Steele, N., "Engaging Our Audience through Photo Stories," *Visual anthropology,* 2005.

[13] McDonald, D.W., "Evaluating Expertise Recommendations," *Proc. Group 2001,* 2001, 214-223.

[14] Moran, T.P. and Carroll, J.M. (Eds.), *Design Rationale: Concepts, Techniques, and Use.* Erlbaum, 1996.

[15] Orlikowski, W., "Learning from Notes: Organizational Issues in Groupware Implementation," *Proc. CSCW '92,* 1992, 362-369.

[16] Shirky, C., "Ontology is Overrated: Categories, Links, and Tags," 26 May 2005. http://www.shirky.com/writings/ontology_overrated.html

[17] Stylos, J. and Myers, B.A., "How Programmers Use Internet Resources to Aid Programming," submitted for publication, 2005.

[18] Trec Tracks, 9 Feb. 2005, http://trec.nist.gov/tracks.html.

[19] Udell, J., "Publishing a project weblog," Jon Udell's Weblog, 27 March 27 2003. http://weblog.infoworld.com/udell/2003/03/27.html

[20] Udell, J., "The weblog as a project management tool," *Tangled in the threads,* 2001.3, 24 May 2005. http://udell.roninhouse.com/bytecols/2001-05-24.html