Comparing Tagging Vocabularies among Four Enterprise Tag-Based Services

Michael J. Muller
IBM Research / Collaborative User Experience
One Rogers Street
Cambridge, MA, USA 02142
+1 617 693 4235
michael_muller@us.ibm.com

ABSTRACT
We compare four tagging-based enterprise services, which respectively stored bookmarks to webpages and documents, to people, to blog entries, and to hierarchically-structured activity records. Analysis of user data and tag data showed relatively small overlaps in tags used. Conventional normalization strategies produced only modest improvement. These results suggest difficulties in combining exploratory searches across multiple social-tagging services. We recommend strategies for cross-service tag integration at the points of tag storage and tag search, rather than at the conventional point of tag entry. We close with a research agenda around this strategy.

Categories and Subject Descriptors
H.5.3. Group and organizational interfaces/Collaborative computing & CSCW.

General Terms
Measurement, Experimentation, Human Factors,

Keywords
Social software, tagging, tag-suggestion, auto-completion, corpus

1. INTRODUCTION
In our research on social computing within the enterprise, we have been interested to understand how people interpret and use social tagging services in different contexts (see [14, 16, 19] for general reviews). This paper provides the first look at tag usage across multiple tagging services.

1.1 Social Tagging on the Internet
Social tagging occurs when a user associates a resource (such as a URL) with a user-chosen descriptor word or phrase (called a tag) in a shared environment in which other users can search and view the user’s tags. The first and most well-known form of social tagging occurred in the shared bookmarks of the website del.icio.us (for the first study of these practices, see [14]). In this service, users created bookmarks similar to those used in individual’s web browsers, usually with the addition of (a) one or more tags and (b) additional free-form descriptive text. The user could search her/his own collection of bookmarks, or could search the bookmarks that had been created by other users. Soon, the concept of tagging was extended to shared photographs (www.flickr.com), shared videos (www.youtube.com), shared evaluations of movies (www.movielens.umn.edu -- see also [33]), blogs (www.technorati.com), and an increasing diversity of more specialized resources such as books (www.amazon.com), books in one’s own library (www.librarything.com), and persons (www.facebook.com). Several experimental services have experimented with online collaborative games to expand the set of tags on selected resources (e.g., [35]).

These services have in common the concept of a resource with a unique online representation, which is tagged by one or more users in an asynchronous shared environment. These services differ from one another in the type of resource being tagged, but also, significantly, in the structure of the tagging record. As mentioned, the original form of social tagging involved an enhanced version of a browser bookmark (e.g., [14, 20, 21]). This structure preserves the user ID of the tagger, and allows the user (tagger) to return and modify the bookmark if needed. Because this structure associates each user with one or more bookmarks and one or more tags, is supports some aspects of social networking [16, 20], as well as the possibility of identifying potential domain experts by the frequency with which they use certain tags [10, 11, 17].

By contrast, some services omit the intermediary structure of the bookmark, or make it unavailable to for editing by the user, and thereby apply the tag directly to the resource (e.g., [10] – see also [34]). This structure tends to reduce the association of user with tag (or user with resource). This structure also provides for a simpler database and potentially increased computing efficiency.

1.2 Social Tagging within the Enterprise
We have been prototyping and studying social tagging systems within IBM Corporation. Tagging services within an enterprise can take advantage of full authentication of users, facilitating the identification of experts [10, 17], people with shared interests [20, 21], features of a shared vocabulary [7], and colleague-based strategies of information search [17]. Some of these strategies may be available, perhaps to a lesser extent, for searches on the public internet.

Full authentication also makes it possible to compare users’ tagging practices in multiple tagging-based services, because
users authenticate in the same manner, with the same user ID, in all enterprise services. This paper goes beyond the previous research, by examining tagging behaviors across four distinct tagging-based services, applying to four types of resources, with partially-overlapping user populations, within a single enterprise. This paper takes a comparative view that was not possible in previously published single-service analyses. The comparative view provides insights into the strengths and weaknesses of tag-based storage and tag-based search across services, and suggests new approaches to solving problems of an integrated view across tag-based services.

Golder and Huberman [14] described a pattern of vocabulary stabilization with increasing tag usage on a particular resource (i.e., across multiple bookmarks that were created by multiple users to describe that resource). Similar results were later obtained by Damianos et al. [7] and Sen et al. [33]. This robust finding suggests that groups of users tend to establish a reliable and consistent set of descriptors for a particular resource, and presumably a consistent usage pattern for each tag. Each of these studies examined a different social-tagging service, and each of these studies restricted its analysis to the data within that respective single social-tagging service. We hope to learn if people expect each tag to have a stable, consistent meaning across multiple social-tagging services, and if their tag usage is consistent across the multiple types of objects stored in the multiple services. Clark’s concept of common ground suggests that people need concepts and words in common in order to engage in collective actions (e.g., [6]). Tagging services in general appear to offer a means for achieving such common ground. Tagging services within a work-oriented enterprise would seem to be a particularly promising setting for people to engage in the co-construction of their common understandings. Our studies will help us to understand the semantics of social-tagging services, and thus the extent to which these services can provide a set of common descriptors for an enterprise’s collective knowledge.

2. Four Enterprise Tag-Based Services

We collected tag data from the four services summarized in Table 1. Examples of the structure of tag entries are given in Table 2.

### 2.1 Dogear

Dogear is a social-tagging service for resources such as public URLs, company-internal URLs, and other company internal documents (e.g., Wiki pages, Domino documents, etc.) [21]. Dogear bookmarks are readable by any authenticated user in the company. An example of a Dogear bookmark could be a URL to a sourceforge project, with a bookmark title of “List Manager”, a brief comment about the significance of the content, and “tags” such as “ajax”, “UI”, “list-management”, and “sourceforge”.

Dogear became operational on 12 July 2005, and had accumulated a total of 21072 tags entered by 1710 users on more than 72677 URLs and documents, as of the closing date of this study. Dogear continues to be used for daily work [20, 21].

### 2.2 Bluepages+1

Bluepages+1 (BP+1) is an enhanced version of the company’s online Bluepages employee directory [10]. Among its enhancements is the ability for one person to apply a tag directly to another person’s directory page. Bluepages+1 data (including tags) are readable by any authenticated user in the company. An
example of a BP+1 tag could be the name of an employee, tagged with her/his role (e.g., “projectmanager”), hints about expertise (e.g., “large systems”, “federal”), and personal information (“football”, “numismatics”).

Bluepages+1 became operational on 16 November 2005 as an extension of the set of experimental Fringe services [9]. Bluepages+1 had accumulated a total of 2992 unique tags entered by 731 users describing 7601 people, as of the closing date of this study. BP+1 continues to be used for daily work.

2.3 Blog Central
Blog Central is an internal blogging service, open to any employee. All entries in Blog Central are readable by any authenticated user in the company. An example of a BlogCentral tagged record could be the blog entry and author-supplied tags such as “GROUP2005report” and “conference”.

Version 1 of Blog Central became operational during October 2003. Blog Central had accumulated a total of 3322 unique tags entered by 2092 users to describe 23143 blog entries, as of the closing date of this study. Blog Central continues to be used for daily work.

The Blog Central data structures provide for a separate list of tags for each blog and for each entry within each blog. There is no automatic propagation of tag data from one Blog Central record to another. All tags must be entered manually.

2.4 Activities
Activities [22] is a web-based version of ActivityExplorer [13, 27], an activity-centric collaboration service in which teams may create a collections of diverse objects in a tree-like structure consisting of a root “activity” and its daughter components. Unlike the other three services, Activities restricts access to each activity to a specified access control list, which may be as small as one person, or as large as several hundred people. An example of an activity with tags could be the structured collection of objects (documents, messages, chats, etc.) called “Write GROUP 2007 paper” with tags such as “conference”, “GROUP2007”, “social-tagging”, and “Deadline:21May”, and with membership of the co-authors of the paper and selected internal reviewers.

Activities became operational on 23 November 2005. The Activities data structures provide a separate list of tags for each object. As of the close of the data collection for this study, there was no automatic propagation of tags from Activities record to another. All tags were entered manually, for a total of 3269 unique tags entered by 742 users to describe a total of 10290 activities and activity components. Since the close of the study, automatic tag inheritance has been added to the Activities system, such that any child object automatically receives (at the time of creation) any tags that had been associated with its parent object. The Activities server continues to be used for daily work.

2.5 Services Used for Real Work
As summarized in Table 1 and shown in Figure 1, each of the services has been used over a long period, and users have accumulated a large number of tagged resources in each service. Dogear has become a standard means of sharing internal and external webpages and addressable documents [21]. BP+1 has been used not only for reputation management ([10]; see also [11, 25]), but also for organizing lists of colleagues related to both expertise and project involvement. Blog Central is the primary point of informal publication within the company. Activities [22] is used by several hundred people every week to coordinate intra- and inter- organizational work, surpassing the kinds of uses originally anticipated for an activity-centric service (e.g., [27]).

The data in Table 1 and Figure 1 reflect only the addition of new resources to the services. The action of adding tags to a previously-entered resource does not appear in those presentations, and the use of the services by readers is also not visible (see [20] for an analysis of search and readership in Dogear). These additional categories of use, while substantial, are beyond the scope of this report.

This theme of real work and frequent access will be revisited below in the section on “Discussion of Low Tag Re-Use Rates.”

3. Method
3.1 Data Sources and Initial Treatments
For Dogear and BP+1, it was possible to obtain, respectively, a log containing one record for each bookmark as it was entered. These logs were collected and preserved on 21 July 2006. Each record provided full information about the date, time, tag, user, and resource. Each log record also contained additional data that were specific to each service; these additional data will not be analyzed in this paper.

For Activities, there was no similar log of bookmark entries, so we obtained a copy of the DB2 database tables from that production server, as of 21 July 2006. Information pertaining to the contents of activities was considered confidential and sensitive, and was kept within a secure server which could be accessed by fewer than 15 people. Only the less sensitive data of date, time, tag, user, and activity-node-ID were extracted for the analyses in this paper. Additional data, which will not be reported here, will in future allow an analysis of tag re-use among the parent/child relationships within the hierarchical activity structures (see [13, 22, 27] for details about these hierarchical structures).

There was a delay in obtaining a similar copy of DB2 tables from the Blog Central production server. A copy was finally saved into a highly secure server on 9 August 2006. As with the Activities data, all contents of entries were kept within the secure server environment, and only the less sensitive data of date, time, tag, user, and blog-item-ID were extracted for the analyses in this
Additional data, which will not be reported here, will in future allow an analysis of tag re-use among the blog-root and blog-node (or blog entry) hierarchical structures.

### 3.2 Data Reduction and Representation

Data from all four services were reduced to a “normal form” consisting of date, time, tag, user, and resource ID. The type of resource was of course different for the different services, so the resource-ID data were used only to differentiate items within each service. By contrast, dates, times, tags, and users were comparable across services, and were analyzed where appropriate on a cross-service basis.

Where possible, data from the people responsible for a particular system were removed from the analyses, in case their usage was somehow unrepresentative, as was done in [11, 20, 21]. Very small teams were responsible for the Dogear and BP+1 prototypes, and it was easy remove their data prior to analysis (4955 records or 2% of Dogear records from three users and 384 records or 2% of BP+1 records from four users). However, because Activities was a product, many people had had some form of time-limited relationship with the Activities server. It was not possible to make a simple determination of a small number of team members whose data could be deleted before analysis, so we analyzed all records from the Activities service. Similarly, because Blog Central was a large internal project that ran two different versions of a production server, many people had some

<table>
<thead>
<tr>
<th>Service</th>
<th>Dogear</th>
<th>Blue-pages+1</th>
<th>Blog-Central</th>
<th>Activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>u=1710</td>
<td>-</td>
<td>293</td>
<td>359</td>
<td>613</td>
</tr>
<tr>
<td>t=21072</td>
<td>-</td>
<td>1196</td>
<td>2953</td>
<td>1694</td>
</tr>
<tr>
<td>t=2992</td>
<td>293</td>
<td>-</td>
<td>193</td>
<td>228</td>
</tr>
<tr>
<td>t=7166</td>
<td>1196</td>
<td>-</td>
<td>900</td>
<td>522</td>
</tr>
<tr>
<td>t=7166</td>
<td>2953</td>
<td>900</td>
<td>-</td>
<td>1100</td>
</tr>
<tr>
<td>t=3269</td>
<td>1694</td>
<td>522</td>
<td>1100</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 3A. Users-in-common in the six pairwise comparisons of services.

<table>
<thead>
<tr>
<th>Service</th>
<th>Dogear</th>
<th>Blue-pages+1</th>
<th>Blog-Central</th>
<th>Activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>u=1710</td>
<td>-</td>
<td>293</td>
<td>359</td>
<td>613</td>
</tr>
<tr>
<td>u=713</td>
<td>293</td>
<td>-</td>
<td>193</td>
<td>228</td>
</tr>
<tr>
<td>u=2092</td>
<td>359</td>
<td>193</td>
<td>-</td>
<td>289</td>
</tr>
<tr>
<td>u=742</td>
<td>613</td>
<td>228</td>
<td>289</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 3B. Summary mean users-in-common across different numbers of services.

<table>
<thead>
<tr>
<th>Number of Services</th>
<th>Mean users-in-common</th>
<th>Mean % users-in-common</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 (six pairs)</td>
<td>329.17</td>
<td>41%</td>
</tr>
<tr>
<td>3 (four triples)</td>
<td>137.5</td>
<td>19%</td>
</tr>
<tr>
<td>4 (one quadruple)</td>
<td>79</td>
<td>11%</td>
</tr>
</tbody>
</table>

### Table 4A. Tags-in-common in the six pairwise comparisons of services.

<table>
<thead>
<tr>
<th>Service</th>
<th>Dogear</th>
<th>Blue-pages+1</th>
<th>Blog-Central</th>
<th>Activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>t=21072</td>
<td>-</td>
<td>1196</td>
<td>2953</td>
<td>1694</td>
</tr>
<tr>
<td>t=2992</td>
<td>1196</td>
<td>-</td>
<td>900</td>
<td>522</td>
</tr>
<tr>
<td>t=7166</td>
<td>2953</td>
<td>900</td>
<td>-</td>
<td>1100</td>
</tr>
<tr>
<td>t=3269</td>
<td>1694</td>
<td>522</td>
<td>1100</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 4B. Summary mean tags-in-common across different numbers of services.

<table>
<thead>
<tr>
<th>Number of Services</th>
<th>Mean tags-in-common</th>
<th>Mean % tags-in-common</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 (six pairs)</td>
<td>1394.17</td>
<td>36%</td>
</tr>
<tr>
<td>3 (four triples)</td>
<td>672.00</td>
<td>22%</td>
</tr>
<tr>
<td>4 (one quadruple)</td>
<td>396</td>
<td>13%</td>
</tr>
</tbody>
</table>

### Table 5. Summary of example of calculating other-similarity and self-similarity of two corpora.

1. Divide each of two corpora (CorpusA and CorpusB) respectively into three equal-sized fractions.
2. Iteratively calculate the similarity between pairs of fractions.

Note: For simplicity, this example used three equal-sized fractions per corpus, whereas in practice we used a decile analysis of ten equal-sized fractions, for a total of 90 fraction-to-fraction comparisons in each analysis.

<table>
<thead>
<tr>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>*</td>
<td>*</td>
<td>B1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>*</td>
<td>*</td>
<td>B2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>*</td>
<td>*</td>
<td>*</td>
<td>B3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mean Self-Similarity = \[ \frac{[ S(A1,A2)+S(A1,A3)+S(A2,A3)+S(A2,A1)+S(A3,A1)+S(A3,A2) ]}{6} \]

Mean Other-Similarity = \[ \frac{[ S(A1,B2)+S(A1,B3)+S(B2,B3)+S(B1,B2)+S(B1,B3)+S(B3,B1) ]}{6} \]
form of time-limitation relationship with one or both versions of the Blog Central service. It was not possible to make a simple determination of a small number of team members whose data could be deleted before analysis, so we analyzed all the data from the Blog Central service.

Several of the systems contained logged data from their very first day of beta or alpha operation, even though there were bugs in the data-logging routines. In consequence, some records contained missing or obviously incorrect data (e.g., dates that were earlier than the first date of service). We rejected the following numbers of normal-form records from each service for these reasons: Dogear: 14545 records (5%); BP+1: 0 records (0%); Blog Central: 361 records (1%); Activities: 0 records (0%). Because each system was has been used for real work by a large population of users, these rejected records constituted only a minimal percentage of our total dataset: 14906 records in across all four datasets, or 4% of all data records.

4. Analysis of Tags

For the purposes of this paper, data consisted primarily of user-tag tuples. Note that previous studies of tagging behaviors within a single service have included not only user and tag (as in our study), but also the tagged object [7, 10, 14, 17, 21]. By contrast, this paper is concerned with tagging across diverse services, where each service is concerned with different types of tagged objects (e.g., documents vs. people vs. blogs vs. activity components). The same objects cannot appear in different services, because each service stores a different class of objects. Therefore, for this initial report, we omit references to the tagged objects, and focus only on the users and the tags.

4.1 Users in Common

Analyses showed complex patterns of overlaps (Table 3). Many of the 4987 unique users participated in more than one of the services. On a pairwise basis, the number of people writing tags into each respective pair of services ranged from 193 users (BP+1 and Blog Central) to 613 users (Dogear and Activities), with a mean pairwise users-in-common figure of 329.17, or a mean users-in-common percentage of 41% (calculated per [4] as the Overlap Coefficient – i.e., by dividing the users-in-common by the smaller of the numbers of users in each of the two source services). Smaller numbers of users participated in more complex intersections of the services, with only 79 (11%) of the users participating in all four services.

The mean of 329.17 users in common across pairs of services provides ample opportunity to explore the potential for common vocabularies across services.

4.2 Service Vocabularies: Tags in Common

When we examined tags in the four services, we found similar, complex patterns (Table 4). Of the 28460 unique tags, many appeared in more than one service. On a pairwise basis, the number of tags appearing in each respective pair of services ranged from 522 tags (BP+1 & Activities) to 2953 tags (Dogear & Blog Central), with a mean two-service tags-in-common figure of 1394.17, or a mean tags-in-common percentage of 36% (calculated, as above, by dividing the tags-in-common by the smaller of the numbers of tags in each of the two source services). Smaller numbers of tags appeared in more complex intersections of the services, with only 396 (13%) tags appearing in all four services.

4.3 Service Vocabularies: Self-Similarity vs. Other Similarity

These are surprisingly small numbers of tags-in-common, if we consider that users were engaging in real work – and the same real work – with all four systems (Figure 1). But how large an overlap should we have anticipated, in view of the well-known phenomenon of users choosing different words to describe the same computational concept [12]?

To answer this question, we adapted a method from the comparison of text corpora. Kilgarriff [18] argued that measures of similarity or dissimilarity between two corpora required a baseline or standard of comparison. He developed statistical methods that could be applied in identical ways, and with identical power, to ask (a) how similar were two different corpora, and (b) how self-similar was each of the corpora to itself? He used each corpus’s self-similarity as a baseline for evaluating similarity or dissimilarity when comparing one corpus with the other.

The core of Kilgarriff’s method is to split each corpus into equal-sized fractions, and then to perform iterative similarity comparisons among all possible pairs of those fractions. For example, suppose that there are two corpora, corpusA and corpusB. Divide each corpus into three equal-sized fractions. This method could be used as follows to calculate both the similarity between corpora and the self-similarity within each corpus, as follows:

- Other-similarity (between corpora): One fraction in each pair is drawn from CorpusA, and one fraction is drawn from Corpus B.
- Self-similarity (within each corpus): Both fractions in each pair are drawn (without duplication) from the same corpus.

Table 5 provides a summary of this simple example. The same respective pairs of fractions are used to calculate the self-similarity of CorpusA, the self-similarity of CorpusB, and the other-similarity of CorpusA vs. CorpusB. It remains to specify the correct number of equal-sized fractions, and the method used to calculate similarity for each pair of fractions.

For each service, we took all the tags1, randomized their order, and then divided them into ten equal-sized fractions, or deciles. To calculate the self-similarity measure of each service, we performed 90 similarity calculations, one for each pairwise set of deciles, omitting identical deciles which would of course have had perfect self-similarity (i.e., in Table 5, there are no comparisons along the major diagonal of each corpus’s matrix, or no comparisons of A1 with A1, A2 with A2, and so on). The similarity measure was simply the ratio of the number of tags-in-common across the two deciles, divided by the mean size of the deciles. Thus, perfect similarity would produce a ratio of 1.0, and zero similarity would produce a ratio of 0.0.

1 Previous analyses were based on the unique tags in each service. We used all the tags in this analysis, because a restriction to unique tags would have led to self-similarity figures of zero.
To calculate the other-similarity measure for two services, we performed 90 similarity calculations, one for each pairwise set of deciles, but in this case we drew one fraction from the first service, and one fraction from the second service. For strict comparability, we omitted the analogous ten pairwise fractions to those that were omitted in the self-similarity calculations (i.e., the major diagonal), resulting in 90 pairs that were analogous to the 90 pairs in the self-similarity calculations.

As a result of this process, for each pair of services, we had (a) 90 measures of other-similarity between the services, and (b) two sets of 90 measures of self-similarity within each service. We calculated the median of each of these 90 sets of fractional similarity measures, resulting in six other-similarity summary measures (one for each pair of services), and four self-similarity summary measure (one for each service). These medians are displayed in Figure 2.

As Figure 2 shows, all self-similarities are stronger than all other-similarities. The calculation of sampling variance in these kinds of overlapping analyses is difficult, but we can use a simple Mann Whitney U test to show that the four self-similarities are significantly greater than the six other-similarities (U=0, p<.01). By the logic of Kilgarriff’s approach, this is strong evidence that there is greater tags-in-common within a service than there is between any two services. This more detailed approach corroborates the conclusion of the earlier section (“Service Vocabularies: Tags in Common”) that there is relatively low commonality of tags across services.

4.4 Service Vocabularies: Relative Frequencies
Tags occurred with different relative frequencies in different services. Using the Fisher r-to-z transformations, we calculated the mean correlation of relative tag frequencies across services as r=.286. While significant (p<.001), this correlation explains only 8% of the variance. People appear to use tags in somewhat different ways in the different services.

4.5 Personal Vocabularies
We were surprised at the low commonality of tags across services. Therefore, in addition to comparing lists of users and tags on a service-vocabulary basis (above), we also conducted analyses at the level of each person (personal vocabulary) and each tag (tag membership).

When we focused on personal re-use of tags across pairs of services, we found very low tags-in-common rates (mean of 1.06 tags-in-common per person, or mean of 2.65% of per-user opportunities to have tags in common). When we re-focused our analysis on the tags, and asked whether each tag was associated with the same users across services (a relationship that we called “tag membership”), we again found very low rates of commonality (mean of 0.35 people associated with the same tag in different services, or mean of 4.89% of per-tag opportunities to have users in common).

We considered that these results might have been due to a sampling oversight, i.e., the mistake of including people who might have created only a very small number of tags in a service. To test this sampling-error hypothesis, we sorted each set of data in terms of the minimum number of tags created by the user in each of the services under analysis. We then experimented with temporarily removing the users with the lowest numbers of tags in one service or the other. We systematically tried all possible cut-off usage rates. None of the cut-off rates substantially increased the tags-in-common on a user-by-user basis. None of the cut-off rates substantially increased the number of people associated with each tag. The phenomena of low tags-in-common and low tag-membership appear to be robust across users with any number of tags.
4.6 Discussion of Low Tags-in-Common
It could be argued that these results are not very surprising: If people were describing different types of objects, then shouldn’t they use different vocabularies to do so?

This explanation is not supported by the data. If people used different vocabularies for different types of objects, then we should have found very few tags in common between services. In fact, we found a minimum of 522 tags in common across pairs of services. Thus, the re-use of the tags, on a service-by-service basis, provided ample opportunity for the re-use of tags on a person-by-person basis. Nonetheless, we found very little personal re-use of tags.

We also note that, at a conceptual basis, all four services were designed to support people’s daily business work. Furthermore, people in fact do use those services for real and daily work, as briefly summarized above in the section “Services Used for Real Work” (see also [10, 11, 13, 20, 21, 22, 23]). On this basis, it appears reasonable to expect people to use similar vocabularies to describe the related aspects of their work that involved different types of objects in different tag-based services. That is, we could easily expect that a user would bookmark a website (in Dogear) using the same project-related tags that s/he used to bookmark a collaboration partner (in BP+1) and that s/he used to describe that project in a posting in Blog Central or Activities. But this did not appear to take place very often, as reflected in the low re-use of tags on a person-by-person basis.

Another possibility is that people did re-use their tags, but did so with variations in spelling or capitalization or punctuation. We explored the following normalization strategies applied to all of the tags from all four of the services:

- “Folding” all alphabetic characters to lower-case
- Stemming — both Porter stemming ([30]; see also [36]) and Paice/Husk stemming [28, 29]
- De-prefixing (e.g., removing a prefixed symbol such as “-” or “%”)

The best results (achieved with a combination of de-prefixing and Porter stemming) led to a 21.3% improvement. This is promising, but a manual inspection of the most-frequently occurring tags in each combination of services showed many additional missed opportunities to find tagging commonalities. There were three major classes of missed relationships that could not be addressed through the three normalization strategies listed above:

- Examples of missed semantic-level overlaps included the ability to treat the embedded string “blue” as a special company-reference when it occurred within project or product names (e.g., “bluepages” vs. “bigblue” vs. “kcblue” etc.), as well as other significant embedded strings that were not detected and integrated when they occurred as substrings within longer tag names (e.g., “ibm” and “IBM”; the acronyms “rss” and “RSS”; and the word “design”). Note that some of these substrings could in principle be specified generically or by market segment (e.g., “rss,” “blog,” etc.), while others would require installation-specific customization (e.g., “blue” for IBM). Note also that these kinds of lists are not necessarily static: A sales organization or a competitive intelligence department might need to perform frequent updates as its list of client companies or competitors changed.

- Examples of missed person-level overlaps included different spellings or conventions for the same name (e.g., “Kim” vs. “KimChen” vs. “Chen, Kim” vs. kimch@us.ibm.com). Note that these problems arose not in the logged data about user IDs, but in tag data into which a user had written the name of her/himself or another user. This is a common method for recommending a bookmark to another person (see [8]).

- Examples of missed concept-level overlaps included different references for product names and projects (e.g., for an enterprise instant-messaging product, “Sametime75” vs. “st75” vs. “st7.5” vs. “buddylist” vs. “buddy” vs. “cbp2006 instant-messaging [customer-reference]”).

Other research within two of these services has shown that people devote considerable effort into writing tags [21], that people use the tags to search for relevant items across services [26], and that in some cases people use the tags as part of managing their reputations within the enterprise [10, 25]. We are left with the apparent paradox that people work hard to write tags, but seem not to do this consistently for related aspects of their work that are stored in different services.

5. Implications for Initial Tag Entry and Tag-Based Search
It appears reasonable for a user to expect to re-use tags from one service to another. Our results suggest that simple tag re-use may prove more difficult than anticipated. Indeed, our work on an experimental aggregation service based on tags showed that, for some users, sparse tag-based search results were a problem [26].

We consider three distinct strategies for supporting tag re-use: support during the task of entering a tag into a bookmark; support during the storage of a tag into a tagging-service database; and support during the task of conducting a tag-based search.

5.1 Supporting Initial Tag-Entry
Various methods are under investigation at internet and intranet sites to assist users during the task of initial tag-entry (i.e., during the creation of a bookmark) to re-use the tags that are already available at that site. Some sites offer a list of popular tags; other sites provide type-ahead or auto-complete suggestions as the user is beginning to enter the first few letters of a tag; still other sites remind a user of the tags s/he recently used (for review, see [16, 33]). Most of these strategies, if applied to our situation of four distinct services, are likely to prove ineffective for the following reasons:

Popular Tags. Popular tags are unlikely to help users who are in the “thin” part of the “long tail” distribution of tag usage (e.g., [1]; see also [3, 15]), at which the greatest differentiation is provided by low-frequency (high information value) tags. In fact, the over-use of popular tags has been shown to decrease their information value [5], and may prove detrimental to providing a strong vocabulary that can help users with particular queries, with

2 This is a partial list, to protect some sensitive project names.

3 This is a fictitious user name, to protect employees’ privacy.
particularly adverse impact upon “minority disciplines” and communities of practice [24].

**Type-Ahead Suggestions.** Type-ahead suggestions are useful only for a subset of potential overlaps, namely those that have the same root. For example, type-ahead suggestions are relatively ineffective for synonyms that do not occur in dictionary-order proximity. In fact, type-ahead suggestions are substantially similar to stemming strategies discussed above, and fail to answer the needs of many semantic-level and person-level overlaps. And in practice, the largest of the four enterprise systems in this paper already implements a form of type-ahead-based assistance for tag-entry [21], and its vocabulary continues to contain many variants on what appear to be common tags [22]. One possible compromise would be to provide the user with only her or his own tags, to improve self-consistency.

**Tag Database Evolution.** At a conceptual level, all of these strategies may be flawed because each operates at a single moment in time, whereas a tag database is constantly evolving with new entries. Helping a user to enter an appropriate tag or tags at one moment, tells us little about the appropriate associations with tags that are added to the database at later moments. It is, of course, possible for a user to edit a bookmark and update the tags associated with it at a later date, but this occurs relatively rarely in practice. Thus, the best we can hope for, with tag-entry support, is to integrate a bookmark with the current state of the database, but not with future states of the database.4 While this point may seem obvious, it reveals a weakness of the tag-entry strategies, as contrasted with other kinds of interventions to strengthen tag re-use across services.

**Summary:** Weaknesses with any Method Emphasizing a Moment-in-Time. In summary, the weakness of all of the strategies during initial tag-entry is that the best they can do is to integrate with the available tags at a particular moment in time. If there is evolution within a particular tag-based service, then the integration at earlier moments in time provides decreasing help for later moments in time. If there is an addition of a new tag-based service, then there is no historical integration based on previous moments in time. To achieve an integration that works for both present and future states of tag-based services, we need a different strategy.

5.2 Supporting Tag Storage

A second approach involves storing each tag as entered, and storing various processed versions of the tag, using an enhanced set of the methods discussed in “Discussion of Low Tags-in-Common,” above. For example, a tag that was entered as “OrganizationalLearning” would be stored with that exact text, but would also be stored as “organizational-learning” (“folded” to lowercase), “organizational-learning” (normalized punctuation), “organiz-learn” (stemmed via Porter stemming, with normalized punctuation), and possibly “organiza-learn” (stemmed via Paice/Husk stemming, with normalized punctuation).

Additional enhancements could include lexical lookups (e.g., synonyms) and a vector of frequently co-occurring tags. Xu et al. [37] and Halpin et al. [15] proposed such methods for finding related tags within a single service (see also [2]). It should be possible to extend that method to recommend related tags in a multiple-service environment, based on the co-occurrence of tags within each component service.

Social recommendation provides another potential strategy for calculating related tags. This approach begins with one or more user-provided tags, and suggests other tags that were used by either (a) other people who used the same tags; (b) other people whose overall tagging patterns were similar to that of the current user; or (c) other people who are closely related to the user in a social network analysis. Similar strategies are frequently used in recommendation systems (e.g., [32, 33]). It should be possible to extend that method by applying tags used by related persons within one service, to make tag suggestions in a multiple-service environment.

In general, the strategy of supporting tag storage has the advantage of preserving any special spelling or punctuation conventions of the user [23], while also providing a more standardized representation that may be able to span the kinds of idiosyncrasies that we encountered in this study. A second advantage of this strategy is that it makes relatively few real-time performance demands. The various methods for processing and normalizing tags could be conducted overnight or at other moments of low performance demand.

5.3 Supporting Tag-Based Search

A variation of the preceding strategy could be adopted at the point of search specification, rather than of tag-entry or tag storage. At a high level, this strategy calls for (a) providing the means for a user to indicate the names of tags to be searched, and then (b) suggesting or recommending additional tags that the user-optional be added to the search. The step of suggesting could use data from a single tag-based service, or from multiple tag-based services.

This strategy would allow for tag re-use at the moment of search, and could make use of one or more of the normalization approaches described in the preceding sections. Additional approaches that leverage the availability of multiple tagged databases could also be included. Unlike the strategy of supporting tag storage, the search strategy can make use of the current state and context of one or more databases. As a result, the search-time strategy avoids the problems outlined in “Tag Database Evolution,” above. However, a disadvantage of the search-time strategy is that it increases real-time computing demands.

6. Conclusion

This paper began with descriptions of four tag-based enterprise services which have supported users’ real work for over a year. The paper then explored the extent of overlap (tags-in-common) among the tag vocabularies in those four systems, finding surprisingly little overlap, especially at the level of personal vocabularies. Based on those findings, the paper critiqued
strategies that encourage tag normalization at the time of initial tag-entry (e.g., during the creation of a bookmark), and recommended instead that mitigation of tag vocabularies be focused at the points of tag-storage and of tag-based search.

The strategies of assisting at the moment of storage or the moment-of-search have the additional benefit of allowing new, experimental methods to be prototyped once and added to the search menu as a cross-service integration feature, without requiring any changes within multiple participating services. Thus, these strategies may provide a useful and flexible testbed for additional re-use concepts. Such a testbed can support a research agenda based around converging comparisons, such as

- How do users choose among the various types of search supports?
- How do users integrate those search supports with their working practices?
- Which types of search-supports provide straightforward mental models for users?
- How do users’ tagging behaviors change as a result of using certain search supports (e.g., [33])?
- How do users’ information-sharing practices change as a result of using certain search supports? How do their social networks change (e.g., [5, 20])?
- How well can communities of practice and minority disciplines make use of selected tags to organize their own vocabularies, resources, and communities (e.g., [24])?
- What indices of search effectiveness can be used to measure the different types of search supports (e.g., [5])? Is it possible to develop a generic, behaviorally-based measure of search effectiveness that does not require explicit user feedback or evaluation (e.g., [32])?
- How can the internet tagging-based games of [35] be translated into the enterprise, such that they deliver value to individual users and to their organizations [8]?

We anticipate combining these analyses with other examinations of tagging patterns across diverse services and resources, to understand the emerging tag-based work practices, and to clarify the value of social software data and services within the enterprise.

7. References


