

# Impact on Performance and Process by a Social Annotation System: A Social Reading Experiment

Les Nelson, Gregorio Convertino, Peter Pirolli, Lichan Hong,  
and Ed H. Chi

Palo Alto Research Center  
3333 Coyote Hill Road  
Palo Alto, CA 94304  
{lnelson, convertino, pirolli, hong, echi}@parc.com

**Abstract.** Social annotation systems such as SparTag.us and del.icio.us have been designed to encourage individual reading and marking behaviors that, when shared, accumulate to build collective knowledge spaces. Prior work reported on the experimental design and performance effects observed in a controlled study of SparTag.us. Study participants working independently on a sensemaking task who had access to a set of expert annotations were compared against participants using SparTag.us without those annotations and participants using only office software for annotation support. A learning effect favored the participants exposed to expert annotations. In this paper, we analyze the behavioral data captured during the experiment and identify differences in the work process that can explain the performance effects reported previously.

**Keywords:** Convergent measures, social annotation systems, evaluation, social sensemaking.

## 1 Introduction and Approach

Learning and knowledge handoff are becoming critical in the workplace. In fact, knowledge work is increasingly depended on (or equivalent to) learning and professional development (Tapscott 1996). Also, corporations are often losing critical domain knowledge as older workers are leaving before they could transfer their knowledge.

But new technologies are showing new opportunities. Web 2.0 tools have lowered the costs for social construction of knowledge (e.g., Wikipedia, social bookmarking) and made possible user-defined combinations of content across web services (e.g., mashups). New semantic web techniques are also allowing users to give structure to the content that they share (e.g., microformats in blogs).

In this context of new social needs and social technology, researchers in Human-Computer Interaction and Information Retrieval are redirecting their focus of inquiry from solitary individuals working with systems and content to models of social information foraging, knowledge sharing, and sensemaking [7, 11]. Social annotation

systems such as SparTag.us [2] and del.icio.us<sup>1</sup> exemplify socially constructed understandings of content.

However, the settings in which collaborative software is employed are full of experimental confounds: real-world socio-technical systems introduce greater complexity into the evaluation process [2]. Measures of performance in social sensemaking remain difficult [4]. Researchers of communication and collaboration technologies have faced this problem repeatedly. As a result, they have started including also process measures (e.g., measures of costs in the process such as turn-based measures of efficiency in communication). Monk and collaborators [5] reconstruct the evolution of measures in these studies. They observe that the traditional measures of task performance such as number of errors and completion time were only sensitive to gross changes in the technology utilized. For example, when performing tasks within experiments, the participants may tend to protect their primary task and get the work done efficiently through extra effort (i.e., costs) at the expense of any secondary tasks. In order to capture these hidden effects the researchers have introduced measures that characterize aspects of the process of communication, rather than just the final outcomes (e.g., number of the turns, length of the turns, kind of turns, see [9]). More recently, experiments on knowledge sharing in teams have combined both process and performance measures in order to assess the effects of the new tool on the sharing process (quality and costs) and then the consequences of these effects on the group performance [e.g., 1].

Previously we reported the main results obtained from performance measures on the use of the social annotation tool, SparTag.us [6]. This show a statistically significant increase in subject matter learning for participants using the tool in a condition of access to annotations of another. In this paper we briefly summarize the method and the effects on performance and then focus on the process measures that were also taken during the experiment. The goal is to use behavioral measures (e.g., URLs visited) and supplemental products (e.g., responses to essay questions) to help characterize visible changes in the process that led to the differences in performance (questionnaire-based outcome). The analysis of these process measures is in part quantitative and in part qualitative.

We next briefly introduce the object of study, the social annotation tool, SparTag.us. We then summarize the experiment and its measures and detail selected process measures found. We discuss these in light of the significant performance gain seen in subject matter learning. We conclude with implications for design and future research.

## 2 The Study System, SparTag.us

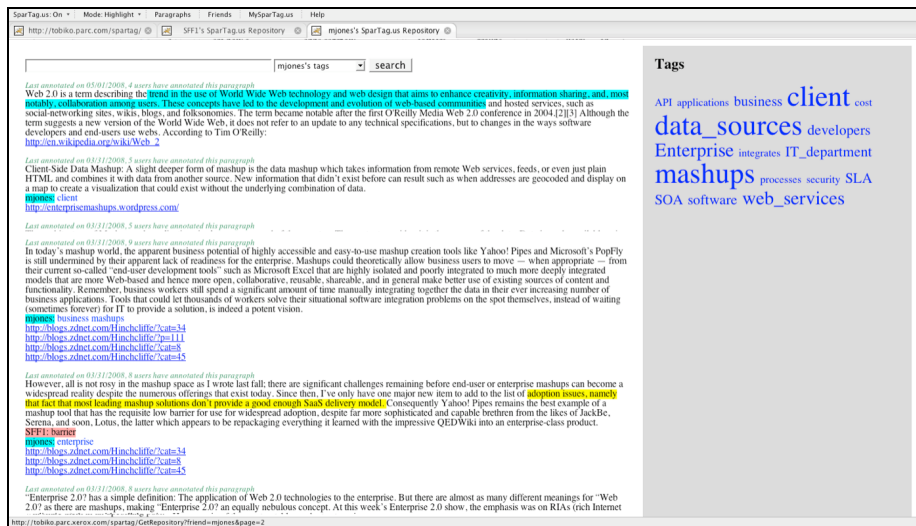
Inspired by the work of Schraefel et al. [8] showing that in many cases of information foraging the content of interest is at the sub-document level, SparTag.us uses annotation as a means to collect paragraphs of interest. Specifically, when a user loads a web page in his browser, we modify the underlying representation of the page to partition its textual content into paragraphs and make the words of the paragraphs live and

---

<sup>1</sup> <http://del.icio.us>

clickable [3]. Here the user can annotate the content in various ways. S/he can click on words of a paragraph to tag the paragraph, This Click2Tag interface offers a low-cost option for the user to annotate paragraphs of interest in situ while reading the web page. S/he can also highlight phrases and sentences in situ through click-and-drag actions. Thirdly, as the user tags or highlights, SparTag.us automatically extracts the annotated paragraphs from the page and inserts them into a system-created notebook, where further annotations can be made later.

In SparTag.us, a user can also subscribe to the annotations of another user by designating that user as a friend. Consequently, the user will see his friend's annotations when viewing pages containing the same paragraphs annotated by his friend. Color-coding is used to distinguish between own and friends' annotations. Figure 1 shows the friend's notebook as viewed by the user. Note that the friend's highlights and tags are displayed in light blue and the user's own highlights in yellow.



**Fig. 1.** Study participants using the SparTag.us annotation tool may view the collected tags and highlights of another in a friend's Notebook. The view here shows annotated paragraph with another's annotations shown in blue and the other person's tag cloud.

### 3 A Social Reading Experiment

We conducted a 'Social Reading Experiment' where participants needed to use Web resources to learn about a topic area: "Enterprise 2.0 Mashups", which is a combination of the technology areas of "Enterprise 2.0"<sup>2</sup> and "Web 2.0 Mashups". Study participants would need to find and understand many web pages because at the time of the study there was no single source of information on the topic area.

<sup>2</sup> [http://en.wikipedia.org/wiki/Enterprise\\_2.0](http://en.wikipedia.org/wiki/Enterprise_2.0)

Our experiment compared three groups of participants who worked:

1. Without SparTag.us (WS), but with traditional note-taking tools.
2. With SparTag.us only, used individually (SO).
3. With SparTag.us with the annotations of a 'Friend' (SF).

The conditions WS and SO were control conditions in which individuals read web content without access to others' annotations. To provide for an ecologically valid comparison, WS participants could take notes in MS Word or with pen and paper. In the SF condition, people independently read web content but also had access to social annotations created by an experimenter-simulated subject-matter expert.

Tools like SparTag.us and del.icio.us are tools used at an Internet scale and scope. In our experimental setup we look at the performance of individual users. However, we extended the scope of inquiry beyond the individual by simulating a social reading condition. That is, in one of the conditions each user was exposed to the SparTag.us Friend, which is an organized collection of annotations comprising a tag cloud, a list of URLs, and a set of paragraphs. These annotations are derived from the following social resources. Twenty tags associated with the top 100 annotated URLs from a del.icio.us query on "enterprise mashup" constituted the target tag cloud. URLs found by top hits from a Google search that used each tag as a search term. These URLs were manually tagged with these 'expert' tags using SparTag.us.

The hypothesis is that participants that were exposed to tags, URLs, and highlights from a knowledgeable other would perform better than the participants without this exposure. We thus evaluate performance measures between subjects in the experimental condition, SF, with those in control conditions, SO and WS. Eighteen participants completed two experimental sessions. The first day was a four hour series of demographic survey, true-false question answering, learning in the domain area lasting two hours, one writing essay, and a debrief. Day 2 lasted one hour and involved one true-false question set and a second writing task. More details on the procedure can be found in [6].

We used a combination of performance and process measures to understand the impact of the annotation support used, but also give indications of how people are employing the technology in the context of their reading and annotation practices. The performance was measured using a questionnaire (created for this study). The questionnaire included a set of true-false questions, which were generated from an expert elicitation process and were used to assess objective learning gains in the subject matter domain before and after the users foraged the information in each of the three conditions.

The process measures pertained to the reading and writing behaviors of each participant: the number and sequence of Web resources visited (logged by Universal Resource Locator or URL), loaded and scrolled; the annotations made (tags and keywords used), and the personal notes taken during the task.

The main measure of learning (equation (1)) was obtained through a metric of learning effect developed as part of the experimental method. The Gain metric is a composite indicator that was computed on the basis of several scores derived from the questionnaire: Pretest to Posttest questionnaire scores for each participant, and maximum score. Specifically, gain scores were calculated as:

$$Gain = \frac{(PostTest\_Score - PreTest\_Score)}{(Max\_Score - PreTest\_Score)} \quad (1)$$

Using the Gain metric as the measure of learning performance, we report in [6] a learning effect, with the SF group showing significantly greater gains than the SO group and the WS group. The WS and SO groups were not significantly different.

This establishes that participants with access to resources from a knowledgeable other exhibited a greater learning performance. What might have caused this effect? What costs are reduced or what kinds of benefits are increased? In the remainder of this paper we turn to the other measures taken during this experiment to explore these questions.

#### 4 Results of Process Measures of Reading Activities

We previously reported in [6] differences in reading activities amongst the groups. While not statistically significant there was a consistent trend seen that on average SF participants visited fewer URLs, but spent more reading time on those they visited. This suggests a more in-depth analysis of the fewer sources of information chosen by SF participants.

**Table 1.** Trends suggest different reading behaviors between conditions

Group	URL Visits		Time on URL	
	Mean	SD	Mean (sec)	SD
SF	59	23.7	144.5	73.0
SO	71.2	25.5	128.2	56.1
WS	79.3	35.9	87.3	24.0

We look further into this by examining what kind of URLs were being visited. Sites were classified as representing the following kind of information sources:

- Blog, indicating the site was an individual's Web log;
- Conference, an industry or academic conference site;
- Consultant, the business site of a consulting service;
- Employment, a job posting site;
- MySpartagus, use of the SparTag.us Notebook;
- News, a general or technology news service;
- OpenSource, information site of the Open Source community;
- Search, an Internet Search service;
- Vendor, a site of a business selling in the domain area;
- Wikipedia.

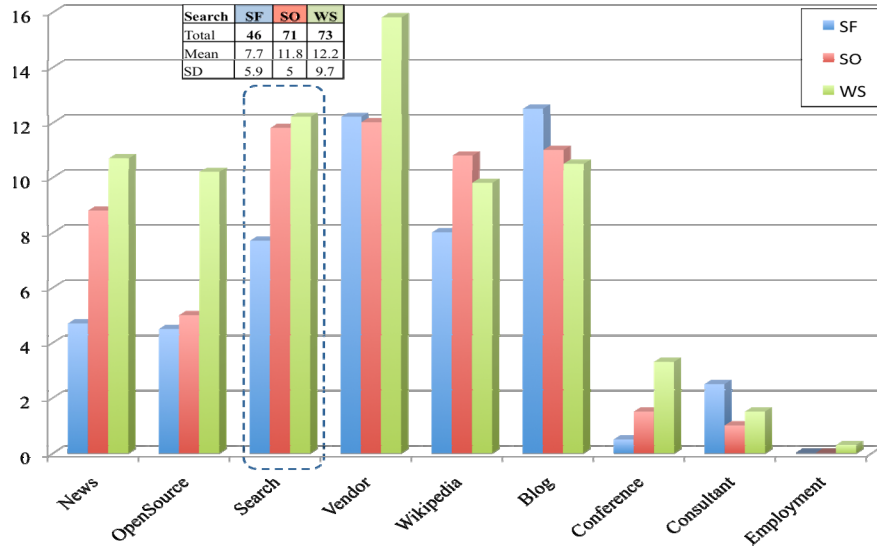


Fig. 2. Counts of URLs visits by kind of source while Reading. Search actions are fewer for SF than for SO and WS participants.

Our aim here is to look at the differences in Web resources used to discover traces of cost/benefit differences between the study groups. Figure 1 shows the classification of the 1149 web sites visited by the 18 participants during the reading portions of the experiment. We can see that search is a main resource (16.5% of all URL visited) and that search behavior exhibits the same trend of distinguishing SF against SO and WS in terms of lower use of search during the observed information foraging.

### 5 Results of Process Measures of Writing Activities

We previously reported [6] differences in writing activities amongst the groups. While not statistically significant there was a consistent trend seen that on average SF participants used more domain terminology in writing answers to their essay questions (see Table 2).

Table 2. People using SparTag.us used more domain words

Group	All Words	SD	Domain Words	SD
SF	549.92	207.01	141.50	46.27
SO	528.92	202.08	136.00	58.68
WS	459.67	174.32	117.00	48.17

Similar to the previous section we used a classification of the kind of URLs visited during writing activities. We looked at only new Web resources that were used during the writing tasks to supplement those found during the prior reading activity. This includes far fewer URLs visited, where ‘Consultant’ Web sites were not visited and new kinds were seen (i.e., private library catalogs and pages from online copies of published books). Table 3 shows the collected use of new Web resources during writing. We again see a trend indicating need for less information finding amongst SF participants over the other groups.

**Table 3.** Participants in the SF condition accessed fewer supplemental Web resources to answer the essay questions

Group	Total	M	SD
SF	39	6.5	5.1
SO	58	9.7	7.7
WS	60	10.0	12.6

## 6 The Devil Is in the (Process) Details

We had found that users supported by our social annotation tool and having access to annotations from a domain expert (i.e., SF condition) showed a significant increment in subject matter learning. In this study we addressed the question of ‘how’ (i.e., ‘in what ways’) such improvement in performance had occurred. To this end, we analyzed process measures characterizing the foraging behavior both before and during the writing of the report (i.e., number of URLs read, average time spent per URL, kind of URL, number of searches). We found that compared to the other two conditions the participants in the SF condition exhibited greater efficiency in foraging (i.e., fewer sources visited, more time per source, fewer searches) and greater efficiency in producing the final report (i.e., more words written in the same with less additional foraging activity done while writing).

Given the experimental differences imposed among the conditions (WS, SO, SF) and the abovementioned better learning performance of the SF participants, these results about the process suggest that having access to the annotations from a domain expert reduced the costs of foraging information, promoted more focus and depth of analysis, and saved time that the SF participants used to write content in the report.

These results point to directions for future work. A more detailed explanation of the how and why these effects occurred will help us understand how they could be induced in other situations (e.g., in peer-to-peer group collaboration or within communities of practice). More detailed exploration of the stimuli used in the interventions is needed.

The SO condition (with SparTag.us only) affords collecting relevant paragraphs in a notebook and, for each paragraph, highlighting relevant sentences, or labeling it with any of its own word (click-to-tag) or with a new user-entered keyword. The SF condition was, by design, the condition with highest support for learning because it further included structured information including three kinds of stimuli:

1. A cloud of tags that represented the expert's terms for the domain.
2. A set of URL to jumpstart the foraging process from relevant sources.
3. A set of sample paragraphs in the expert's (or friend's) notebook which were examples of pieces of relevant information that the user could expect to find and then annotate in her/his notebook.

As part of our future work we plan to examine in detail which of these three sets of stimuli has more effects, at what stage of the process, and potential interactions between them. This requires focused follow-up studies that adopt measures consistent with the present study. These could manipulate solely the exposure to the expert's tags and then measure the effects on the terms typed while foraging new information; or could manipulate only the exposure to the expert's set of URLs and measure the volume and ordering of the sources read and annotated in the user's notebook; or, finally, could manipulate the visibility of sample paragraphs in the friend's Notebook and measure if the final reports by users exposed are individually more focused on fewer topics and/or more consistent among each other (e.g., measure, within each condition, how similar the content of the report is to the content of the paragraphs and/or the cloud tags). In summary, what beneficial effects do experts' tags, URLs, and relevant paragraphs have? How and at what stage of the process are learners influenced by these expert traces? What are the extra cognitive costs when these cues are missing?

## Acknowledgments

Research funded by the ONR grant No. N00014-08-C-0029 to Peter Pirolli. We thank Christoph Held and Diane Schiano for their assistance in designing and conducting this study.

## References

1. Convertino, G., Mentis, H.M., Rosson, M.B., Carroll, J.M., Slavkovic, A., Ganoë, C.H.: Articulating common ground in cooperative work: content and process. In: Proc. CHI 2008, pp. 1637–1646. ACM, New York (2008)
2. Grudin, J.: Groupware and social dynamics: Eight challenges for developers. *CACM* 37(1), 92–105 (1994)
3. Hong, L., Chi, E.H., Budiu, R., Pirolli, P., Nelson, L.: SparTag.us: A low cost tagging system for foraging of web content. In: Proc. AVI 2008, pp. 65–72. ACM, New York (2008)
4. Kalnikaite, V., Whittaker, S.: Social summarization: Does social feedback improve access to speech data? In: Proc. CSCW 2008. ACM, New York (to appear, 2008)
5. Monk, A., McCarthy, J., Watts, L., Daly-Jones, O.: Measures of Process. In: CSCW requirements and evaluation, ch. 9, pp. 125–139. Springer, Berlin (1996)
6. Nelson, L., Held, C., Pirolli, P.L., Hong, L., Schiano, D.J., Chi, E.H.: With a Little Help from My Friends: Examining the Impact of Social Annotations in Sensemaking Tasks. In: Proc. CHI 2009 (2009)
7. Pirolli, P.: A Multilevel Science of Social Information Foraging and Sensemaking. In: Position Paper, Information Seeking Support Systems Workshop, Sponsored by the US National Science Foundation, Chapel Hill, NC USA (June 2008)

8. Schraefel, M., Zhu, Y., Modjeska, D., Wigdor, D., Zhao, S.: Hunter Gatherer: Interaction Support for the Creation and Management of Within-Web-Page Collections. In: Proc. WWW 2002, pp. 172–181 (2002)
9. Sellen, A.J.: Remote conversations: the effects of mediating talk with technology. *Human-Computer Interaction* 10, 401–444 (1995)
10. Tapscott, D.: *The Digital Economy: Promise and Peril in the Age of Networked Intelligence*. McGraw-Hill, New York (1996)
11. Vicente, K.J.: HCI in the global knowledge-based economy: designing to support worker adaptation. *ACM Transactions on Computer-Human Interaction* 7(2), 263–280 (2000)