Analyzing the Role of Tags as Lightweight Traceability Links

Matthew Hale	Noah Jorgenson	Rose Gamble
Dept. of Computer Science	Dept. of Computer Science	Dept. of Computer Science
University of Tulsa	University of Tulsa	University of Tulsa
Tulsa, OK USA	Tulsa, OK USA	Tulsa, OK USA
matt-hale@utulsa.edu	noah-jorgenson@utulsa.edu	gamble@utulsa.edu

ABSTRACT

Tagging offers a traceability mechanism for software development by connecting artifacts in a meaningful way. Our integrated courseware, SEREBRO, provides a framework of tools that capture conversation and artifact creation and modification throughout the software development lifecycle by student team members developing non-trivial software products in a Software Engineering course. Using a data driven approach, we investigate the use of lightweight tagging mechanisms applied by student software project teams and present some preliminary results of this investigation.

Categories and Subject Descriptors

D.2.6 [**Software Engineering**]: Programming Environments – *Integrated environments*; D.2.9 [**Software Engineering**]: Management – *Programming Teams;* D.2.7 [**Software Engineering**]: Distribution, Maintenance, and Enhancement – *Documentation.*

General Terms

Documentation, Experimentation, Measurement

Keywords

Tagging, Knowledge Management, Traceability, Collaboration

1. INTRODUCTION

Effective team communication is essential to the success of the software development cycle [1]. Modern web frameworks offer a collection of lightweight tools [2], such as wikis, forums, and task managers, to support discussion, planning, design, documentation, and implementation activities surrounding each phase of the development lifecycle. Individual tools may be compartmentalized or loosely coupled resulting in artifacts that are segregated across a variety of file types (e.g. software architecture documents, use case diagrams, image files, etc), knowledge management (KM) spaces (e.g. wikis) or embedded in communication messages (e.g. forum posts or email).

Traceability is the ability to follow the life of a software artifact from an idea's inception to creation and throughout

modification. It is also a means for modeling the relations among software artifacts [3]. Typically, this requires that all of the related artifacts be linked together explicitly through software documentation. Specifying and updating these links can add significant overhead to the development process and is often lost when busy software teams neglect to forcibly codify it, potentially resulting in poorly documented software systems [4]. When sufficient discussion meta-data is codified, a traditional manual content mapping approach can be used to interpret discussion and link it to artifacts. However, this manual mapping process is time consuming and tedious for Subject Matter Experts (SMEs).

Tagging offers a promising alternative to both forced manual link tracking to achieve traceability and manual explicit SME content mapping to provide ongoing and post project analysis. Previous work [5-7] suggests that social tagging mechanisms can be applied, outside of their typical social media domain, to the software development lifecycle for tracing code to tasks. We investigate the use of tags as content mapping mechanisms to directly link social and development artifacts in our courseware, SEREBRO, used for interaction and project management by student software development teams. Artifacts in SEREBRO can be clustered, searched and visualized by their tags similar to Web 2.0 social tagging mechanisms (e.g. Flicker[1], del.icio.us[2]). Thus, instead of defining traceable connections directly, artifacts are voluntarily and "organically" tagged to semantically group or cluster them into a larger set of related artifacts.

In this paper, we show early results from SEREBRO that suggest tags can be used as content mapping mechanisms and by SMEs to analyze the traceability among software product artifacts developed by student teams. We discuss the implementation of tags in SEREBRO and present our experimental setting. We define our methods of investigation, a set of research questions, and present preliminary tagging results discovered through SEREBRO's information capture mechanisms. Results are based on data collected from undergraduate computer science majors at the University of Tulsa as part of a Senior Software Projects class.

2. RELATED WORK

A large portion of tagging research focuses on the use of social tagging mechanisms [8]. Tags are typically applied to social media such as photos, music, or other objects such as bookmarks

---

[1] **http://www.flickr.com/**

[2] **http://del.icio.us**

accessible by people on the web. These tags link related social media objects to yield semantic classifications based on content.

Treude et al. [6] found that development teams using IBM's Jazz eagerly adopted tagging mechanisms to tag and manage work items. They showed that tags can denote short duration work tasks or may span the entire project duration. Storey et al. [7] developed an eclipse plugin, known as TagSea, for developers to tag code objects directly. These tag types are typically used to mark code to connect it to work items or for programmers to development breakpoints [7].

Research into traceability suggests that, despite its critical role in the production of "maintainable, adaptable and extensible systems" [9], it is often neglected by software teams [10]. Other work by Murta et al. introduces a high level modeling tool that tracks traceability links, once created, and adjusts the linkages between artifacts as the artifacts are modified [11]. Maintaining these linkages is an important part of a traceable documentation and modeling process. If the model no longer describes the system as a result of dated linkages, then project teams will have a much more difficult time adjusting the system in the future.

3. EXPERIMENTAL SETTING

Undergraduate software development teams from the University of Tulsa use SEREBRO as courseware for managing software projects [12]. This capstone course covers an entire academic year with both short and long team projects to immerse students into software engineering concepts through direct development of work products. Tagging usage data for analysis is taken from the SEREBRO data captured during the 2010 Software Projects classes.

SEREBRO provides a framework of tools including a custom Gantt chart and calendar for task management and planning, a Wiki for project documentation, a shared file space for uploading non-version controlled content such as diagrams or documents, and a Subversion (SVN) repository for managing source code changes. Each of these tools allows for development artifacts to be tagged. SEREBRO's core feature is a graph-based

forum known as an *idea network* [13] for synchronous and asynchronous team communication regarding project activities.

3.1 The SEREBRO Idea Network

Figure 1a shows a sample idea network (blue box) in SEREBRO 3.0, which displays a conversation among a team about their prototype web application. Topic discussions begin when someone posts a *brainstorm* node (blue circle). A team member can *agree* (green triangle) with a post to continue the discussion with his/her own ideas, *disagree* (orange inverted triangle) and add a counter argument, or *comment* (talking bubble) with questions or neutral statements. Multiple brainstorms can be used within a single topic to produce independent discussion trees. A post appears when a user hovers over a node. Clicking on a node displays the post and allows the user to respond, edit, or tag its content. Clicking in the red oval in Figure 1a will bring up a tag menu view similar to that shown in Figure 1b to add a tag from the existing set.

3.2 Tagging and Clustering Artifacts

Figure 1b shows a tag menu related to the artifact titled "Presentation" at the bottom of Figure 2, to assert an additional "feasibility" tag. These previously used team tags are displayed for semantic control. A user may select from the list of tags, e.g. the feasibility tag being added, or insert new tags in the field separated by commas. All artifacts, regardless of type or place on the system, are tagged in a similar fashion.

Artifacts can be searched using the Tag Search feature, which returns a *Tag Group* containing all artifacts tagged with the query tag. Users can view all traceable links between the different artifact types, click on them to review their contribution to the overall Tag Group, and sort them to target a certain type of artifact, such as uploads. Figure 2 shows a sample tag search using the tag "presentation" in one of the software team's space. The red box lists the artifact type, the green box lists the name or header of the artifact, the blue box shows the artifact creator (user), the date the artifact was created (date), and the topic location for idea types (thread).
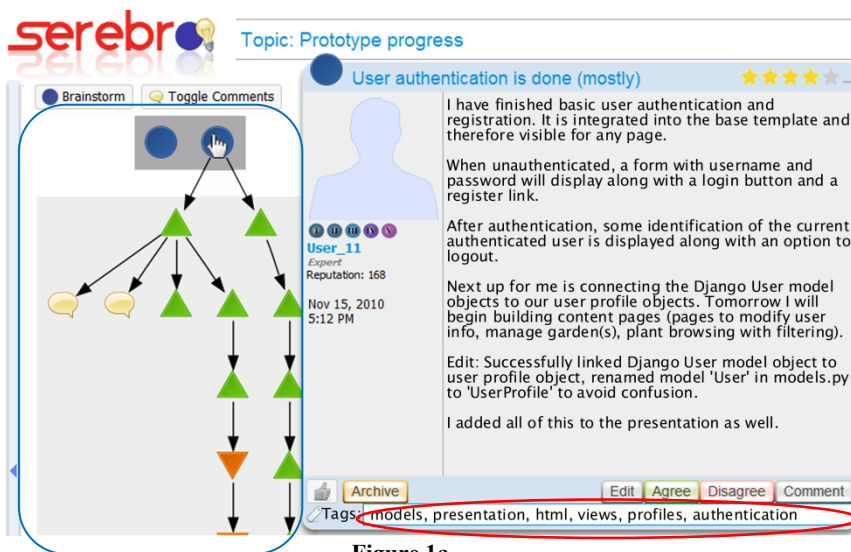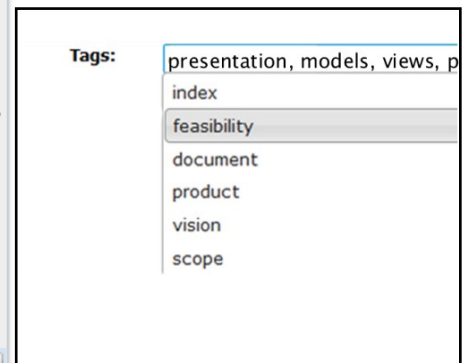

Figure 1a

Figure 1b

Figure 1: (a) SEREBRO Idea Network with post and (b) Applying Tags to Wiki Post

In this example, there are four development artifacts tagged: the tagged wiki entry titled "Presentation" (bottom of Figure 2) and three MSPowerPoint files uploaded to SEREBRO later. We also see that 7 ideas, spanning the creation times of all four artifacts, were tagged with the "presentation" tag which likely means they contain group discussion around the topic. Notice the post from Figure 1a appears as the 3rd entry in the tag search, since it contains the "presentation" tag. The buttons in the upper left allow the artifacts to be sorted by type.

Figure 2: The "presentation" Tag Group

Tag Search lets project teams view and track traceable links through tag application. In addition to *Tag Search*, SME's can analyze traceability using the SME Tag Analysis Module (TAM) shown in Figure 3. With the TAM, SMEs can sort threads by tag (top right Figure 3), selecting only certain posts while simultaneously viewing other linked artifact items, such as those in the bottom of Figure 3. The SME TAM is used for on-going and post project analysis to trace design decisions through discussion and artifact creation.

Figure 3: SME Tag Analysis Module

4. RESEARCH QUESTIONS

We are interested in how tagging mechanisms applied to a set of development artifacts provides traceability across the development lifecycle within SEREBRO. Specifically we explore research questions pertaining to the use of tags for content mapping, SME analysis and team tracking. Our main research question is stated below.

RQ1: Do tags provide traceability through the software development lifecycle and are they a feasible alternative to traditional, time consuming, content mapping methods? For a preliminary answer to RQ1, we start by examining the following:

(a) What percentage of Tag Groups contain at least one idea and one development artifact?
(b) What percentage of development artifacts have traceable links to ideas via tags?
(c) What percentage of ideas have traceable links to development artifacts?
(d) How many total traceable links exist?
(e) How many total traceable links does tagging find?

The results of these questions will guide improvement of SEREBRO's tagging mechanisms to support student software project management. We expect three novel contributions:

- A form of traceability that does not require a large team time commitment to maintain and update.
- Improved tool support to facilitate embedding tags into the software development process.
- Better understanding of the links that exist between communication and development artifacts
- SME support for rapid project analysis of artifacts

5. METHODOLOGY AND EVALUATION

Since SEREBRO 3.0's tagging features were introduced in Fall 2010, we are still collecting data across the academic year. However, we have already seen trends in usage. Preliminary results for RQ1 (a) – (c) and (e) have been calculated using a sample of the ongoing projects, shown in Section 6.

Our proposed methodology is a data driven approach composed of two types of quantitative data. The first type of quantitative data is measured using SEREBRO tagging mechanisms. We examine and extract quantitative data from artifact and idea tagging activity and the resulting Tag Group compositions. The second type of quantitative data is measured directly using traceability artifact mapping, where artifacts are manually examined for traceable links to other artifacts or ideas. Having both types of quantitative data will allow us to validate the consistency - that is how reliable are tags for traceability - and feasibility of tagging mechanisms for use as traceable links. We discuss how this data will be used and our progress in the next sections.

6. EMERGING RESULTS

To date, we have data from two semester milestones (a total of 4 teams, with 8 team projects). The data collected includes Tag Groups that primarily span uploads, wiki entries and ideas, due to limited coding requirements during these milestones. As project code requirements expand throughout the year, we expect to see a much higher presence of tagged SVN commit artifacts in Tag Groups. With the current data, we have preliminary answers to RQ1 (a), (b), (c) and (e). Our results are presented in Table 1.

For RQ1(a), we found that 25% of tags create traceable links between development artifacts and ideas. In Table 1, for RQ1(b), on average, roughly 85% of development artifacts have traceable links to one or more ideas. This result signifies a further review of SVN commits is needed. For RQ1(c), nearly 42% of tagged ideas relate to development artifacts. Finally, Table 1 shows that for RQ1(e), tagging found a total of 1057 traceable links, across all projects and teams. Manual content mapping to show traceability between artifact and ideas for

RQ1(d) is not yet complete. Once this analysis is completed we can compare the results to RQ1(e) to determine how traceable tagging performs compared to traditional tagging. Completion of the academic year will also provide a vast amount of data with which to validate our findings.

Table 1: Preliminary results by question

RQ1: (a) Percentage of Tag Groups having at least one Idea and artifact					
Artifact Type	Gold	Silver	Platinum	Titanium	Average
Uploads	29.5%	7.7%	6.6%	2.7%	11.6%
WikiSnips	36.4%	14.3%	12.4%	11.0%	18.5%
All Artifacts	47.7%	19.8%	15.7%	11.0%	23.5%
RQ1: (b) Percentage of artifacts that have traceable links to ideas					
	Gold	Silver	Platinum	Titanium	Average
Artifact count	46	21	34	17	30
Artifacts without links to ideas	6	1	6	4	4
Percentage of Artifacts with links to ideas	87.0%	95.2%	82.4%	76.5%	85.3%
RQ1: (c) Percentage of ideas that have traceable links to artifacts					
	Gold	Silver	Platinum	Titanium	Average
Total Number of Ideas	263	393	413	269	335
Ideas without links to Artifacts	79	225	232	238	194
Percentage of Ideas with links to	70.0%	42.7%	43.8%	11.5%	42.0%
RQ1: (e) Number of traceable links detected by tagging (shown by type)					
	Gold	Silver	Platinum	Titanium	Total
Uploads	235	102	51	24	412
WikiSnips	234	138	215	39	626
All Artifacts	469	259	266	63	1057

7. FUTURE WORK

We continue to study how SERBRO can leverage Tag Groups and their overlapping references to better collect, maintain and visualize traceable links spread across project artifacts for use in SME analysis and by project teams. Additional research is needed to understand the percentage of artifacts that exist in more than one Tag Group to obtain a picture of how teams strengthen the links among artifacts by associating them with the same Tag Groups. We plan to examine if hierarchies among Tag Groups are naturally established by the team. Much of this research requires manual content mapping to support any automated results obtained by examining the Tag Groups. Similar content mapping is needed to answer RQ1(d).

It is our hypothesis that the availability of tagging mechanisms, (i.e., how easy tagging is for users), has a direct effect on tag usage and thus the resulting traceability of artifacts to ideas. Future student surveys on the use of SEREBRO, including the ease of tagging compared to other social media, will drive the investigation into potentially alternative tagging methods to increase traceability. Finally, we have begun to automate the visualization of traceable links among tag groups for SME analysis. Future work will improve this visualization.

8. CONCLUSION

Our preliminary results suggest that tagging is an effective mechanism for lightweight traceability between upload and wiki artifacts and ideas. Students use the tagging system naturally. They reuse tags across milestones, even when different team members take over a project. While some tags were used that were semantically equivalent, it was not overwhelming. Further examination is needed with regard to SVN artifacts and the complete academic year projects. Given the results of our experimentation, we hope to show that traceability can be achieved with lightweight tags that do not require a large time commitment to maintain and update. Ultimately, we believe that the emergent traceability from the use of the tags will increase the connectivity between the implemented system and the satisfaction of product requirements; that is, that student teams will be able to use SEREBRO's tagging analysis to determine if they have created the right product.

9. REFERENCES

[1] T. Hall, D. Wilson, A. Rainer, and D. Jagielska, "Communication: the neglected technical skill?," *Conf.on Computer Personnel Research*, 2007.

[2] T. Ebadi, M.A Purvis, and M.K. Purvis, "A Collaborative Web-based Issue Based Information System (IBIS) Framework," University of Otago, , New Zealand 2009.

[3] S. Ratanotayanon, S. Sim and D. Raycraft, "Cross-Artifact Traceability Using Lightweight Links," *ICSE Wksp. on Traceability in Emerging Forms of Software Engineering*, 2009.

[4] Aman-ul-haq and M. A. Babar, "Tool support for automating architectural knowledge extraction," *ICSE Wksp. on Sharing & Reusing Architectural Knowledge*, 2009.

[5] C. Treude, "The role of emergent knowledge structures in collaborative software development," 32nd *Int'l Conf. on Software Engineering*, 2010.

[6] C. Treude and M.-A. Storey, "How tagging helps bridge the gap between social and technical aspects in software development," *Int'l Conf. on Software Engineering*, 2009.

[7] M.-A. Storey, L. Cheng, I. Bull, and P. Rigby "Shared waypoints and social tagging to support collaboration in software development," *Conf. on Computer Supported Cooperative Work*, 2006.

[8] M. Ames and M. Naaman, "Why we tag: motivations for annotation in mobile and online media," *Conf. on Human Factors in Computing Systems*, 2007.

[9] H. Gast, "Patterns and traceability in teaching software architecture," *6th Int'l Symposium on Principles and Practice of Programming in Java*, 2008.

[10] Patrick Mäder, O. Gotel, and I. Philippow, "Getting Back to Basics: Promoting the Use of a Traceability Information Model in Practice," *ICSE Wksp. on Traceability in Emerging Forms of Software Engineering,* 2009.

[11] L. Murta, A. Hoek, and C. Werner, "Continuous and automated evolution of architecture-to-implementation traceability links," *Automated Software Engineering.,* vol. 15, pp. 75-107, 2008.

[12] N. Jorgenson, M. Hale, and R. Gamble, "SEREBRO: Facilitating Student Project Team Collaboration," 33rd *Int'l Conf. on Software Engineering*, *Demonstrations track*, 2011.

[13] R. Gamble, *et al.*, "The SEREBRO Project: Fostering Creativity through Collaboration and Rewards," *Education and Multi-Agents Systems Workshop at 8th Int'l Conf. on Autonomous and Multiagent Systems,* 2009.