

Document Labeling, Organizing and Finding in Shared Repositories

Emilee Rader

October 23, 2007

Contents

1	Introduction	1
1.1	Shared Repositories	2
1.2	Proposed Research	4
2	Literature Review	6
2.1	Document Labeling Conventions	6
2.2	Vocabulary Problem and Common Ground	6
2.2.1	Referential Communication Tasks	8
2.2.2	Labeling as <i>Packaging</i>	9
2.3	Organizing	11
2.4	Recognition, Browsing and Navigation	12
2.5	Information Management in Shared Repositories	15
2.6	Concept Map	16
3	Research Plan	17
3.1	Study 1: CTools Project Sites	17
3.2	Study 2A: Labeling and Organizing	22
3.3	Study 2B: Information Management Outcome	25
3.4	Significance and Impact	27
4	Timeline and Proposed Budget	27
4.1	Timeline	27
4.2	Budget	27

1 Introduction

Much of an organization's information is represented in the form of documents, such as reports, memos, meeting minutes, email messages, wiki pages, or blog posts. Ineffective document management incurs costs such as "lost work time, ineffective access to information, duplication of effort, failure to share information, and information overload" (Gordon, 1997). Many different kinds of workgroups including research

terminology
more consistent
in this section
(and
throughout the
document);
increased the
font size

groups, corporate teams, and software developers use shared repositories, storing documents on a server so that others may find and access them. Some examples include group blogs and wikis, and repositories of documents or software code maintained using a diverse set of software tools (e.g.: Drupal, Subversion, Sakai, OpenText Livelink, shared network folders, etc.)

Shared repositories are maintained by many organizations. They are essential for document sharing, and can be greatly beneficial for organizational efficiency, communicating organizational goals, and also for learning and innovation. They can contain “mission critical information” such that if it were lost there would be serious consequences (Blair, 2002). Despite the importance of the information stored within them, shared repositories generally do not have explicit rules or structures for organization and searching, like a library catalog does. Instead, they tend to accumulate content over time and become more and more disorganized, such that users have difficulty finding the documents they need. According to Gordon (1997):

“Why would busy, professional people spend so much time looking for missing documents? Because certain information is *mandatory* for business to be conducted effectively. If a document can’t be located, it can add to the time it takes to complete a task, delay its completion, or prevent it from being completed altogether. A document can encode intense, sustained intellectual activity for which individuals are highly trained and well paid. Such knowledge is part of the backbone of an organization” (p112).

clarification to
problem
description in
the
introduction

Managing group information in a shared repository is a collaborative effort; the information structure of the repository emerges through users’ individual, idiosyncratic labeling and organizing choices and provides the means by which documents are located and accessed. Labeling and organizing documents in a shared repository can be considered packaging the information for later reuse; unfortunately, in most situations, people do not package content effectively for reuse by others (Markus, 2001). The purpose of the research proposed in this document is to improve users’ ability to locate information in shared repositories, by investigating factors that influence users’ choices when labeling and organizing documents.

1.1 Shared Repositories

Shared repositories are online storage spaces used by workgroups for storing and organizing shared documents, and their use is increasing. A shared repository is more complex than just an “aggregate of every individual’s contribution” (Jian and Jeffres, 2006), and maintaining it is a collaborative activity. Documents in a repository are *shared* in the sense that they can be accessed by anyone with permission to use the repository; however, the action of adding a document to a repository is more like making the document available to the users of the repository, than *sharing* it directly with any particular person. Contributing a document to a repository does not guarantee that another user will be aware that it is there, or be able to find it.

clarification of
SHARED
(adjective) vs.
SHARING
(verb); general
revisions for
readability

Several aspects of shared repositories set them apart from other ways information is stored and shared online. A repository user is generally familiar with other repository users through interactions that take place face-to-face or via some communications medium, and with projects and joint work activities they are engaged in together. However, he can expect to be familiar with only some of the documents stored in a shared repository, and he may or may not have been involved with contributing and organizing documents.

Shared repository users can be information *producers*, document authors who create content and contribute it to the repository; and information *consumers*, or re-users of documents in the repository (Markus, 2001). In a personal repository like a laptop hard drive, the producer and consumer are necessarily the same person. However, in a situation where multiple users have access to a shared repository, this is not necessarily true. The producer and consumer roles can be filled by the same individual, or any combination of users, resulting in a situation where a shared repository user might be trying to find documents with which she is unfamiliar or looking for familiar documents stored in unfamiliar places.

In addition, users have different preferences for how information should be organized, which can be a problem for information management in a shared repository (Berlin et al., 1993; Whittaker and Hirschberg, 2001). Unlike libraries or even an organization's website designed by an information architect, shared repositories generally do not have rules for what the information structure should look like; even in instances where rules exist, they are often not strictly enforced. Library records and classification schemes were created for describing content items and codifying relationships between subjects (Rafferty, 2001). In contrast, shared repositories do not necessarily have unified goals or purposes to guide users' choices about how to label and organize. Being *unstructured* means that less effort is required when storing and labeling documents — because there are no rules describing suitable contributions users are free to express what is salient to them about the documents, rather than what might fit within the classification scheme, within the system's constraints (Marlow et al., 2006). However, the lack of structure can also have a down side. Effective packaging for later reuse requires that information producers consider the information needs and context of whomever might want to find and access the document, i.e., the information consumers (Markus, 2001). Producers' unstructured document and folder label choices can affect the future reuse of the information in the repository by information consumers, because decisions about document and labels and locations provide the means by which the documents can be found and accessed. The distinction between those producing the content and those consuming it is an important one: it implies that there is an exchange of information that occurs between users who are producing content, and users who are accessing content, via the repository. This information exchange includes not only the documents, but also the information contained in the document and folder labels and the hierarchy structure. And because document and folder labels are represented in the form of text, factors that shape language choices in communication situations might also affect users' packaging choices.

Furnas et al. (1983) demonstrated that if two random users were to create a label for the same document, it is unlikely that they would choose identical words. Fortunately, users of shared repositories are not necessarily random pairs of people who are unknown to each other. In the best case, they share a work context and even have some knowledge about each other's preferences and personal styles. So, while there is naturally a great deal of variability in users' choices when storing documents in a shared repository, their knowledge about each other and their shared context — their common ground — might mitigate the problem somewhat, if it were somehow brought to bear. Common ground is the mutual knowledge, beliefs and assumptions that people share about each other (Clark, 1996). It is inferred based on joint membership in cultural communities and through shared perceptual experiences, and accumulates via conversation. As conversation progresses, people introduce ideas and vocabulary that become part of their common ground, and can subsequently be referred to without the overhead of having to re-introduce them.

small
clarification of
'feedback'

There is much experimental evidence to support the idea that common ground affects language choices. Speakers tailor their utterances for listeners, with performance implications in experimental tasks (Schober and Clark, 1989). In addition, people create labels for their own use that are different from those created for an unknown future person (Fussell and Krauss, 1989). We tailor what we say to whomever is the intended recipient; it is reasonable to hypothesize that common ground might indeed affect the labels information producers create for documents they store in a shared repository. An important difference between a real-time conversation and any type of asynchronous communication is in the timing of feedback, which is important for establishing common ground and negotiating meaning (Clark and Brennan, 1991). For example, facial expressions are a form of nonverbal feedback that convey whether or not a speaker has been understood by a listener. Shared repository software does not include interface mechanisms that might allow users to provide feedback to each other on document labeling and organizing choices; indeed, what information might be included in that feedback, and what form it might take, are open questions.

paragraph
about goals,
browsing, and
finding

Finally, when users browse a shared repository in search of a document they need, they have a *goal state* in mind that can range from clearly specified (i.e. "I'm looking for a specific document and I know it is here somewhere") to vaguely specified ("I need to find all the information I can related to a past project that I was not involved with"). The representation of the goal state allows the user to recognize when a particular label they encounter while browsing seems related to their desired information management outcome. If common ground affects the labels users choose when creating folders and storing documents, it may also affect browsing by influencing the vocabulary with which users specify their goals.

1.2 Proposed Research

The purpose of this research is to improve users' ability to find documents in shared repository systems, by investigating factors that influence users' choices when label-

added scoping
and
significance &
benefit

ing and organizing documents. This problem is present in a diverse set of software tools for shared repositories; essentially, any time users store documents online for others to find and access. Any given software system makes certain kinds of tasks and goals easier than others, but the underlying fundamental (psychological) problem is the same. People can effectively communicate with each other face-to-face (for the most part) — why is it so difficult to label documents so others can find them?

Information producers are also the consumers, and this characteristic has some interesting and important reflexive implications. In a broadcast format, like TV or online news or RSS feeds, the producers and consumers are different groups of people, and it is easy to separate the distinct roles. But in a shared repository, the choices and behaviors of users when producing or contributing content determine what can be found and accessed in a repository. The document hierarchy is therefore co-constructed and evolves over time; individual behaviors and choices of one user can constrain the options, choices, and behavior of others.

In my dissertation, I ask the following questions about situations in which a group of collocated or distributed users employ a shared repository as part of their work:

1. What patterns exist in the document and folder label choices made by users of a shared repository? What goals and tasks trigger users to look for documents in the repository, and what problems do they encounter?
2. How do the type of common ground (from community membership or shared past experience) and the intended audience affect document and folder label choices?
3. How do the influence of common ground and intended audience on document and folder label choices affect the usefulness of the information structure for finding documents?

I have chosen a mixed-methods approach for my dissertation. Two lab experiments separate the producer and consumer roles, removing the simultaneity of co-construction to assess the impact of common ground on hierarchy formation, and then the impact of different hierarchies on the ability of users to locate documents. In contrast, the field study will allow me to collect data on these aspects of shared repository use in a realistic context and includes data collection about individual behaviors, site characteristics, and group characteristics. The field study will examine a shared repository system in use at the University of Michigan: (ctools.umich.edu). I completed pilot interviews with faculty, staff and students using CTools to support ongoing collaborative projects; I will follow up this pilot with a more focused investigation of users' information management behaviors within their own shared repositories. Two experiments will be conducted concurrently with the field study that will test hypotheses about the effects of common ground on choices made by users when storing, organizing, and finding documents using shared repositories. This research will not only produce implications and advice for designers of shared repository systems; it will also contribute to theory in HCI and psychology regarding language use in an information management context.

clarification in
wording of
research
questions

2 Literature Review

2.1 Document Labeling Conventions

Conventions are spoken or unspoken rules for how people should behave in certain social situations. Such rules, even in distributed collaborative systems, evolve as the system is used (Ackerman, 2000; Krauss and Fussel, 1991). Berlin et al. (1993) encountered problems with conventions when they implemented their own “group memory” system for their research group. Despite agreeing upon conventions for their repository, there were differences in how group members adhered to them. One member of the group commented, “It was hard to remember what we’d agreed to, and what each person remembered tended to drift toward the person’s initial position” (p26).

Sometimes, conventions are agreed to in principle, and then intentionally ignored in practice. Mark and Prinz (1997) conducted a field study of a group using a “large groupware system” to store and share documents. The users of this system held “workshops” in the early days of using the system in order to discuss and decide upon conventions they would follow. After using the system for about six months, it became clear that it was becoming disorganized and unusable, in part because no one was adhering to the conventions. Mark and Prinz (1997) concluded that in this case, it had been too difficult to imagine in advance what conventions would be needed. As the system was used, work practices changed, making the conventions the group had agreed to less appropriate for the situations that arose. Also, there were some users who were unwilling to give up their own, idiosyncratic practices. In some cases it was a conscious choice to violate conventions. One user said, “Naming conventions, reference code, and subject area, I always violate. I give file names that seem to fit” ((Mark and Prinz, 1997); p. 23).

At the outset of using a repository, users don’t know what information structures will work best, and after it has been in use for a while cleaning up the repository is too onerous a task for most users to be willing to undertake (Barreau, 1995). Shared repository systems typically don’t support synchronous interaction among users, nor provide feedback or cues that might communicate and reinforce conventions. Without conventions, it is difficult for information producers and consumers to coordinate their actions with respect to the documents in the repository. For example, if the convention is for meeting minutes to always be stored in one particular folder and someone puts them somewhere else, it could be difficult or impossible for anyone else to find those minutes again.

2.2 Vocabulary Problem and Common Ground

Furnas et al. (1983) reported that random pairs of people use the same label for an object at most 20% of the time. They called this phenomenon the *vocabulary problem*, writing: “There are many names possible for any object, many ways to say the same

thing about it, and many different things to say. Any one person thinks of only one or a few of the possibilities” (p. 1796). Many other researchers have also observed the same pattern (Bates, 1998; Trigg et al., 1999). The implications of these findings for shared repositories are dire: if two random users were to create a label for the same document, they would be far more likely to choose different labels than the same label. Similarly, if an information consumer attempts to imagine what the document he is looking for might be called, chances are low that he will end up looking for the correct label.

Humans’ use of language is imprecise and flexible, and meaning is determined by the surrounding context, and complex communication processes. As a conversation progresses, participants introduce ideas and vocabulary that become part of their common ground, and can subsequently be referred to without the overhead of having to re-introduce them. Common ground is necessary for coordination of conversation, and essential for people to understand one another. Conversation participants believe common ground exists when there is evidence for a “shared basis”. Evidence that a shared basis exists for members of a workgroup using a shared repository can be recognized in the usage of specialized knowledge and language (Clark, 1996). Clark also wrote that conversation participants develop a “feeling of others’ knowing” (p111), a sense of what others do or do not know, that plays a role in assessing how much common ground exists between them. Common ground can be classified into three types (Nickerson, 1999):

- *Shared immediate context* which is ephemeral, existing in the present while two people are in a conversation or working on a task together.
- *Shared past experience*, which is delineated by contemporaneous and collocated past interactions and experiences; i.e., people who have interacted with each other in the past. This type of common ground is created among people who might have taken the same class at the same time and experienced the same events, or worked together on a group project.
- *Community or category membership*, shared by people who have characteristics in common but have never directly interacted. For example, two people who both grew up in Chicago but never met can be said to have community membership common ground. Likewise, two experimental psychologists share this type of common ground, where an experimental psychologist and an accountant would not.

There is some debate in the literature regarding what the cognitive representation of common ground might look like. Clark (1996) does not address the issue of mental representation of common ground; others in the literature have attempted to clarify how a seemingly endless reflexive process might be represented; Lee (2001) calls this the “*mutual knowledge paradox*”. According to Nickerson (1999), common ground can be conceptualized as a “model of others’ knowledge”. He argues that “one’s behavior with respect to others is influenced in various ways by what one knows (i.e., believes, assumes) about what specific others know” (p739). Conceptualizing common ground

added
paragraph
about mental
representation
and common
ground

in this way makes sense from the standpoint of an applied researcher who is more interested in the implications of common ground than the cognitive processes by which it is formed, represented, and used.

2.2.1 Referential Communication Tasks

Referential communication tasks are commonly used in psychology experiments investigating common ground and other aspects of language use in conversation. Schober and Brennan (2003) includes a discussion comparing the use of situated, real-world conversational data vs. referential communication experiments:

“Laboratory studies of task-oriented conversation have the advantage of allowing researchers to assess speakers intentions and addressees comprehension independently of the conversation, through external behaviors like grasping and moving objects.” (p129).

In a typical referential communication experiment, two conversation partners must work together to complete some kind of task. One partner has information the other does not, and they must talk with each other to successfully complete the task. In some instances, the partners are in the same room but not allowed to see each other; in others, communication occurs via a medium such as video or instant messaging. There have also been experiments that used pre-recorded speech in lieu of one of the partners (Chantraine and Hupet, 1994), or involved an overhearer who tried to complete the task without directly interacting with the conversation partners (Schober and Clark, 1989), or a confederate as one of the partners whose role was to introduce certain vocabulary to the interaction (Metzing and Brennan, 2003). Referential communication tasks are often measured in two ways, either by counts of words and speaking turns, representing efficiency, or by evaluating the task outcomes as a measure of effectiveness.

According to Carroll (1980), the three phases that occur in a referential communication task are:

1. trade descriptive phrases (for the object, image, abstract form, etc.)
2. label proposed by one person
3. label accepted by other person and both use it henceforth

Users of a shared repository label shared documents, and engage in activities and tasks which require that they find documents via these labels in order to use them; the labels may have been created by themselves or other people. These characteristics are both present in referential communication tasks. However, there are two crucial differences (Brennan and Clark, 1996; Fussell and Krauss, 1989; Metzing and Brennan, 2003):

- Partners in a referential communication task repeatedly interact with one another, and develop a mental representation or model model of both the history of

extended
common ground
lit review

the interaction and the other person, even when their only experience with each other is for the duration of the experiment.

- Feedback between communication partners in a referential communication task is essential for label agreement to occur. When the interaction history does not exist or the exchange of feedback is prevented from taking place, evidence that an agreement has been reached is not available. This results in communications that are less efficient, and in which miscommunications or mistakes can occur.

Any communication that might take place via a shared repository, if it is even perceived by the users as communication, takes place asynchronously, with little overt transmission of feedback from one person to another. For example, one user may label a document and store it in a particular folder. If another user decides to move it to a different folder, this action may go unnoticed by the first user until they need to find the document again. The reasons for the relocation of the document are unlikely to be communicated via the system. In fact, it may be that users don't even see their usage of a shared repository as a form of communication, because the awareness that others also use the system is not salient to what they are doing at the time. Even if the system had the capability to convey feedback between users, they may not avail themselves of it.

Interestingly, Brennan (1998) used a conversational grounding framework to explain aspects of users' interaction "with and through computers". According to that paper, users naturally expect feedback, even when their conversation partner is a computer:

"Many DOS and UNIX users soon learn to check to see whether their commands have the desired effect; for instance, after copying, moving, or deleting a document, they may list the contents of a directory to discover whether all is as expected. Such checking behavior is a way of grounding with an uncooperative operating system." (p203)

This may mean that if a system were designed with the capability to somehow increase users' awareness of each other and provide avenues for feedback to be conveyed, users might be able to avoid some misunderstandings and coordinate better. A question remains regarding how one might experimentally determine whether various interface designs increase awareness, or whether this might encourage users to provide feedback.

2.2.2 Labeling as *Packaging*

There is much experimental evidence to support the idea that common ground affects language use. Speakers tailor their utterances for listeners, with performance implications. In an experiment conducted by Schober and Clark (1989), participants completed a referential communication task where one participant instructed another how to construct an abstract shape using puzzle pieces. A third participant (the over-hearer') who was not visible to the others and did not speak during the experiment

listened in and tried to construct the same abstract shape with another set containing the same pieces, at the same time. The intended listeners were significantly more accurate at constructing the shapes than the overhearers (98% to 85%). Beliefs about the goals of the listener also affect how speakers construct their utterances. Russell and Schober (1999) found that being correctly informed about a partner's goals had an impact on how much was said and how understanding was displayed. Also, participants assumed others' goals were the same as theirs if they were not told otherwise as part of the experiment.

The previous two experiments both involved synchronous conversation. An experiment conducted by (Fussell and Krauss, 1989) showed that people label things differently for themselves than for an unknown future person. Participants wrote short descriptions of abstract line drawings to help themselves identify the drawings at a later time, or to help someone else identify them. Descriptions were more than twice as long when written for others than for themselves (12.7 versus 5.0 words). When participants returned weeks later, they used the descriptions to identify the drawings. They were correct 86% of the time with their own descriptions, 60% of the time with descriptions written for others, and 49% of the time with descriptions written by other people for themselves. Subjects also had the highest confidence that they had identified the correct shape based on their own descriptions, followed by descriptions written for others, and finally descriptions by others for themselves.

The results of these experiments indicate that common ground might indeed affect the labels information producers create for documents they store in a shared repository. People tailor what they say to whomever is the intended recipient, even when they are simply instructed to write descriptions for "someone else". While a shared repository is not a communications system, language is being used as abbreviations to represent the contents of documents, and also to suggest relationships among groups of documents. Common ground helps us understand each other in conversation; the same might be true when the communication is mediated by a shared repository. Groups with more common ground might assign labels to documents that others in the group will be able to anticipate more frequently than 20% of the time. However, this is not as straightforward as it sounds. (Hertzum and Pejtersen, 2000) wrote:

"Packaging also requires that the professionals suspend their normal way of looking at and working with their documents to take an outsider's look at them. This is, however, difficult because the individual professional has an inherently incomplete sense of whether his/her documents will eventually be of interest to someone else, and, if so, to whom and in what context" (p47).

In other words, simply being aware of others' knowledge, background and joint experiences is insufficient for properly "packaging" information for a shared repository. The ability to take the perspective of others is also necessary. Interestingly, this problem does not occur exclusively in shared repositories. It even occurs between professional catalogers and information seekers. Šaupperl (2004) interviewed 12 catalogers about their process for cataloging, and concluded that they were more concerned about com-

mon ground with other catalogers than with people who might be using the catalog entries they were creating. There are at least three possible perspectives from which the meaning of any given document may be interpreted: the author's, the cataloger's, and the reader's. Šauperl found that the catalogers who participated in the study were aware of this, but mainly tried to stick to the ways similar content had been cataloged by other catalogers in the past, rather than anticipating potential readers' perspectives. According to Šauperl, this seemed to be inherent to the indexing process which requires adherence to structured formats, and that consistency be maintained with the way similar content items have been cataloged in the past.

2.3 Organizing

When storing a document in a shared repository, labeling is only half of the task. Information producers must also decide where the document will be stored, i.e., the folder location in the hierarchy at which someone else will be able to access the document again. Whittaker and Sidner (1996) wrote that filing is a “cognitively difficult task”, because when an information producer is deciding where to put a document, he must imagine where he and others might want to go looking for the document again, as well as remembering how everything else is categorized, the rules and definitions for what each folder contains, and the relationships among the different folders. The consequence for making a wrong choice is that nobody will be able to find the document again. One user said, “I don't know where to put it. And by making a wrong decision, I could really forget about it...” (p. 279). Making this choice gets harder as the repository gets larger, because it is not possible keep all the folders and all the rules in one's head at the same time (Bellotti et al., 2005; Malone, 1983; Whittaker and Hirschberg, 2001). The more folders one has, the less helpful they are at reducing the amount of stuff one has to remember. If each folder contains two or three documents, there are a lot more folders to remember than if each one contains ten or fifteen documents.

Filtering and pruning are activities that information producers typically don't like to do (Markus, 2001), and increases in digital storage space mean that people are able to store more information than ever before. So they defer evaluation, or initially put aside documents that are hard to classify, and only deal with them later if something else happens to prompt action. If this doesn't happen fairly quickly after the document is put aside, it probably won't happen at all (Whittaker and Hirschberg, 2001). People feel like they should hang onto information they aren't sure they need, just in case the need might arise later. Often, later never comes, and people generally don't go back and purge without an incentive or triggering event. Deferring evaluation might mean information producers never get around to thinking about whether a new document should be stored in the repository or not, meaning the documents might not end up in the repository at all.

These findings came from a study of personal information management, but there is no reason to suspect that they would be invalid for shared repositories. In fact, users

small revision:
extended this
example to
include
hoteling office

of a shared repository might be even more reluctant to purge. Imagine a refrigerator in a common area in a workplace. Food accumulates in the refrigerator over time as people forget what they've brought or it gets buried underneath the new arrivals. The older food starts to go bad and get moldy. Eventually, someone just gets disgusted and fed up and starts throwing things away. A similar phenomenon can also occur in other workplace common areas — consider the example of a hoteling office. When several people share an office on a temporary basis, stuff can accumulate just like it does in the refrigerator. However, it is much harder for one person to make an executive decision to throw away other people's stuff when there are no outward signals of spoilage to indicate that the items will no longer be needed. Clearly, nobody will want the moldy pizza; but the choice may not be so black-and-white for a pile of old meeting minutes or out-of-date lab procedures. The occupants of the office must then communicate about the task of cleaning up the office, in person if they happen to run into each other, but more often by leaving notes in the office.

In a shared repository this problem is compounded further because there are not necessarily cues in the interface indicating how recently an item was accessed, and the costs of leaving outdated digital information 'laying around' aren't as immediate as having to push someone's pile of books and papers out of the way to set up your laptop in the hoteling office. In addition, mechanisms rarely exist within a shared repository interface to exchange communicate about a specific document or folder; these communications must be conducted in another software program, or another medium altogether. The path of least resistance is to leave things as-is.

2.4 Recognition, Browsing and Navigation

Recognition memory is triggered by some kind of stimulus or other information in the environment. There are two types of recognition, *familiarity* and *recollection*. According to Yonelinas (2002), familiarity is an automatic, perceptual process — you feel like you've seen something before, but can't remember where; recollection happens when you recognize something you've seen before, and are able to elaborate on that memory once it's been triggered. It is difficult to find studies of the implications of recognition memory processes in real-world situations, rather than lab experiments with little external validity (Elsweiler et al., 2007). The label-following literature in human-computer interaction is one example.

Label-following occurs when users attempt to complete a task using a menu-based interface. Researchers studying label-following were inspired by the cognitive psychology problem-solving literature. The array of choices among possible menu items is the *problem space*, and novice users match the vocabulary they see in the task description with the labels they can see in the interface when choosing what to do next (Polson and Lewis, 1990; Franzke, 1995). A label-following perspective was incorporated into the cognitive walkthrough usability inspection method question, "Will the user notice that the correct action is available?" (Wharton et al., 1994).

In one label-following experiment, Mehlenbacher et al. (1989) hypothesized that user

renamed this section, added lit review about browsing, recognition and label-following

tasks and goals might affect label-following in menus. In the first condition, the task description contained the same vocabulary as the menu items (direct match); in the second condition the task descriptions contained synonyms of the menu vocabulary (synonym); and in the third condition they used pictures and diagrams to communicate what the users were supposed to be trying to do with the software (iconic). Unfortunately, Mehlenbacher et al. (1989) did NOT ask users in the iconic condition to verbalize their interpretation of the pictorial task descriptions. The menu in this study was not hierarchical; it was a flat list, and menu item labels were grouped either alphabetically or functionally. They found that performance was fastest and nearly error-free for the 'direct match' task combined with the alphabetic menu. The functional menu structure was faster for the synonym and iconic task types. There were more errors in the iconic tasks than the synonym tasks; even when they analyzed just trials without errors (for the iconic tasks only), the functional menu was still faster than the alphabetic menu. All conditions showed practice effects such that difference between conditions were nearly eliminated after the first few trials. The results of the Mehlenbacher et al. (1989) experiment indicate that when users are left to generate their own goal specifications, more variability exists in measurable performance outcomes of their label-following behaviors.

In physical space, people make inferences and assumptions about where things “should be” located based on information in the environment. For example, everybody has had the experience of looking for the bathroom in an unfamiliar building — there are places where you just expect to find a bathroom, based on your past experience in other buildings and cues from what you see around you. Information spaces that are arranged in a hierarchical structure have built-in explicit cues about what is located where. Hierarchies may convey information about the structure and content of a shared repository that information consumers would be unable to access if they were to interact with the repository using a search interface only. According to Dourish (2004), “In information work, the meaningfulness of information for people’s work is often encoded in the structures by which that information is organized” (p. 30). Jones et al. (2005) found that folder hierarchies and document labels provide meaningful information that helps people summarize content as well as organize it. Grouping things manually allows for the formation of visible relationships between documents. Visibility into the relationships in an information space might allow an information consumer to orient herself to the content, and choose better where to go next (Chalmers, 2003). It is possible for structure to be inferred from a list of search results and memory for the query that was entered, but this forces the information consumer to work harder to construct structural relationships that can be explicitly stated with a hierarchy (Cutrell et al., 2006).

Chang and Rice (1993) describe browsing as “recognition-based” and “searching without specifying [query terms]”. Information behavior researchers refer to three levels of goal-orientation in browsing (Chang and Rice, 1993):

1. search or directed/goal oriented browsing, i.e. “I’m looking for a specific document and I know it is here somewhere”

2. general purpose, semi-directed browsing, i.e. “I need documents related to a certain project or created by particular person, and I think they might be here”
3. serendipity, undirected, random, not goal-oriented browsing, i.e. “I need to find all the information I can related to a past project that I was not involved with”

When browsing a hierarchical structure, how do people decide where to look next, and when to give up and move on? Pirolli (2005) wrote about information foraging theory, which accounts for and predicts browsing behavior on the web. Information foraging theory states that the links on web pages are “cues” that activate certain cognitive structures related to those cues, via spreading activation. Users will choose to follow links with text that triggers higher activation levels in memory for concepts related to the user’s goal state. Users move on from a given location when the expected potential of the current site (estimated from activation triggered by visible links) is less than that of moving on (estimated from past web surfing experiences). A study by Mobernd and Spyridakis (2007) demonstrated that “navigational link phrasing” — link labeling — affected navigation in a news website; confusing or ambiguous hyperlinks decreased overall comprehension of the information, and discouraged exploration. An experiment conducted by Vaughan and Dillon (2006) found similar results; they created two versions of the same health information website, one which was similar in design and link labeling to typical health information websites, and one which violated users’ expectations for page layout and link labeling. The group that used the “expectation-conforming” version of the website explores more of the site initially than the group using the “expectation-violating” version, but when asked to search for information they were able to find what they were looking for faster.

A shared repository’s information structure is similar in some ways to a website with a link structure: folder labels are like link text. An information consumer is able to browse until she recognizes something related to what she is looking for (Bruce et al., 2004; Trigg et al., 1999). In a study conducted by Boardman and Sasse (2004) users searching their personal repositories used a combination of browsing and sorting of folders. Because they were searching their personal repositories, they exhibited a tendency to know approximately where in the hierarchy to start looking. From there they used recognition memory navigate to the particular document they wanted. Teevan et al. (2004) called this *orienteering*: using recall to make an initial jump to a location from which to start navigating in steps, via recognition, toward the ultimate goal. At each stage, the local context is used to remind people about where they should go for the next step. Teevan et al. (2004) mentioned one participant who tried to find something in her personal repository, but could not explicitly recall the path or any of the folder labels for where it was stored, making it very difficult for her to search for the document using a query interface. Orienteering allowed the participant to find the document, because the information she needed at each step to prompt her next step via recognition was built into the information structure. All she had to do was be able to recognize the next step, not recall it.

However, as I have illustrated previously in this paper one difference between shared and personal repositories is the variance among users in their level of familiarity with

the documents and folders in the repository, and the vocabulary used to label them. Chang and Rice (1993) wrote that goals for browsing can be vague and changing, and users might not know exactly what they are looking for until they see it. It might be that common ground could play a role in how browsing goals are specified; however, it is not clear from the recognition or label-following literatures how users might consider their model of others' knowledge when cognitively matching a vaguely specified goal state with the shared repository structure they encounter, in order to make choices regarding how to proceed.

Finally, Suchman (1994) wrote that hierarchy and categorization serves not only to make things more organized; it can also communicate information about the values of a group. Document labels and the representation of the relationships between documents and people that are made explicit in a hierarchy structure can clearly communicate what, and who, are important and what is not, and reflect power structures within the group. This perspective hints at purposes beyond organizing that hierarchy might serve, communicating information about the structure not just of the information, but of the relationships of the individuals using the information, and the social structures within which they operate.

2.5 Information Management in Shared Repositories

Shared repositories used by small groups or teams are not a known corpus, like one's own documents and folders on a personal computer, nor are they a completely unfamiliar corpus, like a library catalog or the web. This means that some of the documents in a shared repository will be familiar, but most will probably be at least somewhat unfamiliar, and folders may have labels that may seem somewhat misleading or incorrect. This is not likely to be intentional, but as I have shown in previous sections, lack of adherence to conventions, unique individual goals and strategies, and the vocabulary problem make it difficult for information consumers to find what they need in a shared repository. Markus (2001) wrote that information producers tend not to be very good at documenting their work. But she also argued that even when people do a great job at documenting, work "byproducts" like notes and meetings and diagrams etc. can build up to such an extent that too much effort is required to search them:

"For instance, one virtual team committed to using a sophisticated knowledge management system found that they could easily spend 10 minutes out of a 45-minute team meeting searching a 1,000-entry knowledge base for the information they needed. These problems were so severe that team members advocated the use of knowledge intermediaries to help them cope" (p63).

This problem is compounded when the producers and the intended users of the information are not the same people. When information producers document for themselves, they are the beneficiaries of all their hard work. There are few inherent incen-

tives for them to spend time and effort documenting for others; when satisficing, this is likely one of the first tasks to fall off the plate (Greenberg et al., 2003).

A shared repository is a form of external memory, that can “greatly augment what we remember, allowing us to consider and compare much more information than we could keep in our heads. But, more subtly, it can influence how we think as well” (Blair, 2002). Information consumers tailor their information seeking behaviors according to the features and capabilities, or the external representation, with which they interact. A consumer’s information seeking behavior can be expected to be very different depending on whether they are interacting with a query interface, or browsing a document hierarchy.

Lansdale (1988) suggested that personal information management applications for computers should take advantage of the way human memory works, rather than mimicking the ways people manage information in the physical world. Memories are formed as people interpret meaning in a particular context, and the ability to recall details depends on the relationship between how those details are stored in memory, and what is salient about the context in which the person is trying to remember the details. In other words, it is both what we’re thinking about when we store something, and what we’re thinking about when we’re trying to find it, that interact to determine whether or not we’ll be able to achieve success.

Once an information consumer has decided that useful information is likely to be present in the repository, he must browse the repository and make judgments about which documents might contain the information they need. These judgments might be made more difficult in shared repositories than in searching other kinds of information, because contextual information essential to understanding and interpreting the information in the repository is typically not captured with the documents (Hertzum, 1999; Markus, 2001). The process and reasons behind decisions, the “whys” behind the way things turned out, are typically not documented or archived. While a project is active this is may not be much of a problem, because those involved are familiar with the context. But once the project is over that knowledge is rapidly lost (Hertzum and Pejtersen, 2000). Having access to the documents does not mean access to the meaning and implications behind those documents, which were created by particular people in a particular situation for some specific purpose. One must have access to knowledge about the author’s context and purpose to fully understand.

2.6 Concept Map

The concept map shown in Figure 1 on page 18 illustrates the relationships among various factors hypothesized to affect document labeling and organizing in group information management situations, and subsequent information seeking. Three research studies are proposed in this document to investigate these relationships.

The first study is a field study of document labeling and organizing, and information management outcomes in a shared repository, designed to explore the relationship

clarifies the sequencing of the proposed studies

between the existing information structure of a repository (consisting of a hierarchy of labeled folders and documents) and how well it is able to support browsing to locate documents. The field study will also collect general information about how this type of shared repository software (CTools) is used in a real-world setting, because the research literature is fairly sparse in this respect.

A series of experiments conducted concurrently with the field study will focus on common ground and awareness of potential future reuse, and their impact on document and folder labels, and information structures. The mixed-methods approach is important for this research project, because a lab experiment is limited in the degree of external validity it can achieve. The field and lab studies complement each other, by allowing me to gain in-depth knowledge about users' behavior with respect to their own documents and repositories, as well as working towards generalizable principles regarding common ground and language choices. Both the contextualized data from the field study and the findings of the experiments can be applied to the design of future systems; the experiment findings can also contribute to theory in psychology.

- **Study 1 (field study):** What patterns exist in the document and folder label choices made by users of a shared repository? What goals and tasks trigger users to look for documents in the repository, and what problems do they encounter?
- **Study 2A (experiment):** How do the type of common ground (from community membership or shared past experience) and the intended audience affect document and folder label choices?
- **Study 2B (experiment):** How do the influence of common ground and intended audience on document and folder label choices affect the usefulness of the information structure for finding documents?

3 Research Plan

3.1 Study 1: CTools Project Sites

CTools project sites are sites created not for courses, but to support group projects that take place at the University of Michigan. Anyone may create a project site. CTools project sites are primarily used for storing shared documents via the "Resources" section of the site, which presents users with an interface that follows the desktop file-and-folder metaphor for organizing information. I propose conducting a study of information seeking with users of CTools project sites. CTools is an example of a group information management system in use by thousands of people, and it is possible to study their use of the system in the environment in which it occurs. Results will allow me to quantify and describe information seeking activities in greater detail, in the context of the particular system. It is important to investigate people using their own content; artificial information seeking tasks in the lab are less externally valid.

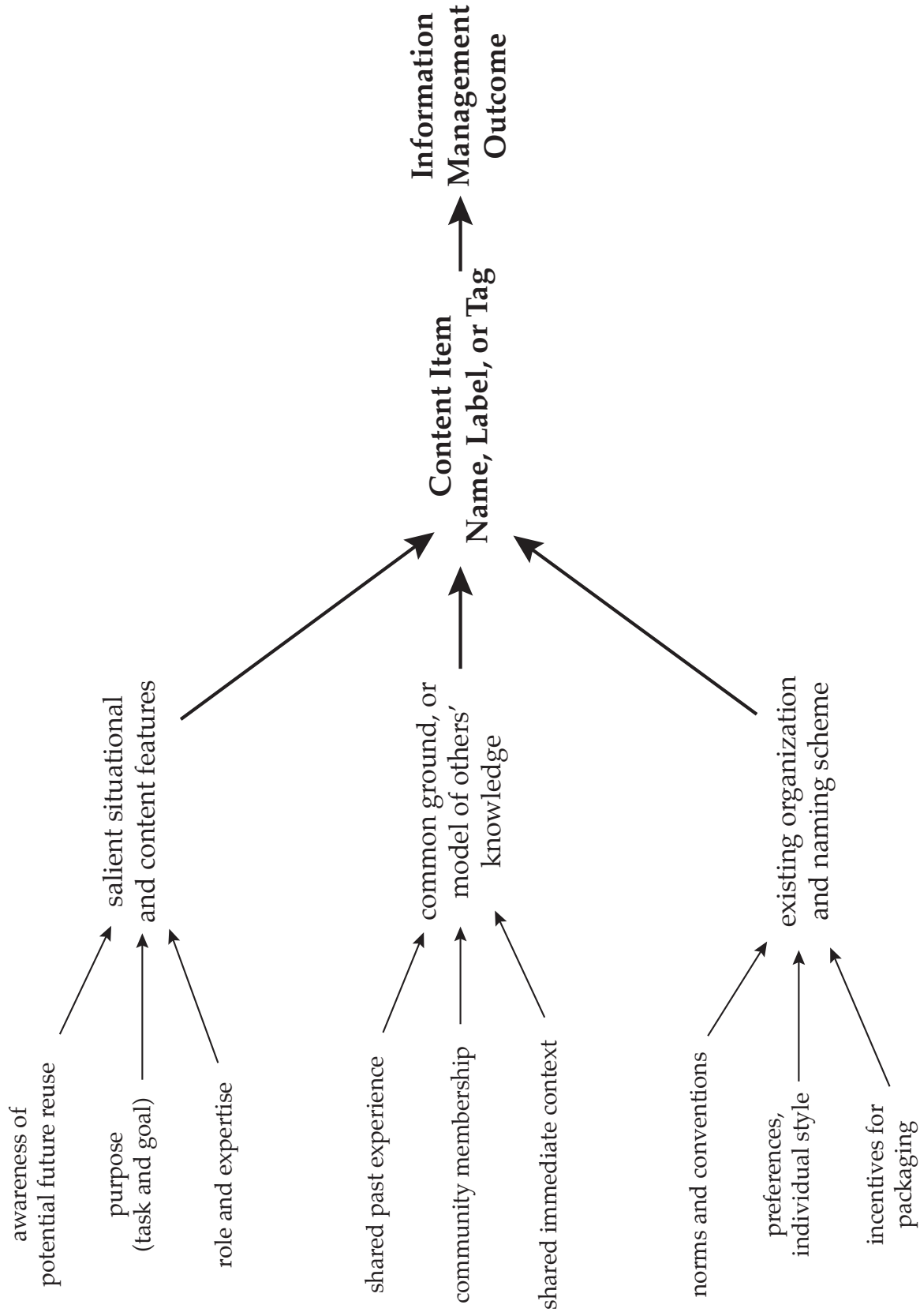


Fig. 1: Concept map depicting hypothesized influences on labeling choices and document finding in shared repositories

Research Question

What patterns exist in the document and folder label choices made by users of a shared repository? What goals and tasks trigger users to look for documents in the repository, and what problems do they encounter?

Sources of Data: Interviews, Search Tasks and Event Logs

CTools project site users will be interviewed and asked to complete search tasks in the context of their own project sites. Search targets will be selected after an initial interview with an “informant” from each project group included in the study. Information collected during an interview with the informant will be used to select search targets along a variety of dimensions.

- frequency and recency of use
- ownership of the item
- depth in the repository (hierarchy level)
- descriptiveness of document label, in relation to the contents of the document

Other members of the same CTools site will be recruited to participate. They will be asked not to use the CTools site on the day of the interview, and at the start of the interview will be asked how long it has been since they used the site. They will be presented with printouts of the search target documents (selected from their own site) one at a time, and asked first to RECALL unprompted where in the site’s Resources folder hierarchy each document might be located, who the owner might be, what it might be called, how old it is, and to describe how it is used. By asking participants first to recall (or guess) information about each document, performance enhancements due to recent exposure to the site are somewhat minimized. After attempting to recall information about each document, participants will be asked to think aloud while browsing their CTools site to find the documents (recognition rather than recall), starting for each document at the top level of the Resources hierarchy with all folders closed. I will counterbalance the order of the search targets by type, according to the dimensions described above, in an attempt to mitigate any order effects.

Informant Interview Questions

- Tell me about the site. What’s the purpose? Who are the group members? What’s your role? How long has it been around?
- Tell me about how you use the site... how often, for what kinds of things? What kinds of documents are in the site? Who are the other primary users?
- Can you give me a tour of the site – talk about each folder and what is in it?
- When was the last time you used the repository before this study session? Tell me about that time – what were you doing – why did you need the document?

Can you find that document for me now (think out loud)? When was the time before that?

- Can you think of a document that you consider important to the group, or used often by the group? Show me. Open the document, tell me about it... Another document? What parts of the site get a lot of use? Not very much use?
- When was the last time you added a document? Tell me about it... can you remember how you went about deciding what to label it and where to put it? Is there something youve been meaning to add? Show me...
- When was the last time you re-labeled, moved, or deleted something? etc.
- When was the last time someone in your group cleaned up the site? Tell me about it... what happened?

Other Site Member Interview Questions

- Tell me about how you use the site... how often, for what kinds of things? What kinds of documents are in the site? Who are the other primary users? What parts do you use?
- When was the last time you used the repository before this study session? Tell me about that time – what were you doing – why did you need the document? Can you find that document for me now (think out loud)? When was the time before that?
- Can you think of a document that is an important document, or used often by the group? Show me. Open the document, tell me about it... Another document? What parts of the site get a lot of use? Not very much use?
- When was the last time you added a document? Tell me about it... can you remember how you went about deciding what to label it and where to put it? Is there something youve been meaning to add? Show me...
- When was the last time you re-labeled, moved, or deleted something? etc.

CTools Event Logs

The interview and search task data will be augmented by an analysis of CTools event logs. I have obtained event logs for January-December 2006, consisting of a record of most users' actions users with CTools project sites during that time period. Any time a user logs in, or creates, views, modifies, or deletes a document located in the Resources section of the site, a record of that action is captured in an event log. All data have been anonymized, but each user was assigned a unique identifier so it is possible for me to segment the data into sessions of activity for a given user and then characterize what types of activities people engage in when using their CTools project sites. It is also possible to follow changes made to individual documents within the

Resources section of a given project site, to try to identify overall patterns in changes to the information structure of the sites over time.

Limitations and Tradeoffs

Because I am planning to collect data with respect to one particular system (CTools), this research should be considered a case study. Each system has a different interface design that supports user needs and goals to varying degrees of success; given unlimited time and resources, it would be better to use the same interview and search task protocol with participants using different group information management software.

There is also a sampling bias. Participants will self-select for this research, meaning that I might end up recruiting people who are more motivated CTools users than average, or are interested in the research topic. It is also important for this research that I interview users and ask them to interact with their own CTools site and their own information, but this makes it difficult to assess how representative the results will be, or even to aggregate findings across people and project sites.

added another
limitation

The document finding tasks in the field study will take place in a compressed time period compared with the timeframe in which these activities would arise as part of a user's normal work context (Mehlenbacher et al., 1989). It may be necessary to adjust the interview plan to include a series of interview sessions in order to space the search tasks over a longer period of time.

Finally, the interview and search tasks investigate the organizing and finding aspects of using CTools project sites, but not the influences on how people choose labels for things. Labeling documents and folders is an intermittent task that happens spontaneously; any staged labeling task for the purpose of this research would be somewhat unrealistic, because the context in which the labeling takes place influences the label that ultimately is created.

Participants

I plan to recruit participants from 8-10 different project sites. In a pilot study of users from CTools project sites, I conceptualized site 'type' in two different ways:

added more
detail on the
types of project
sites

1. Based on the public description of the site entered by the site owner when the site was created
2. Determined by categorizing project sites according to the number of active users, and the frequency of events

I categorized the 50 most active CTools project sites in 2005 according to their public descriptions; the counts were as follows:

- 17 admin (electronic reserves, faculty searches, hr initiative)

- 20 learning (course resources, student course projects)
- 7 research (research activities not associated with a particular course)
- 6 extracurricular (fraternity, sorority, choral group, nonprofit run by students)

I am hoping to recruit from a range of different site types but this will depend to some degree on the responses I receive to my solicitation for participants. I plan to recruit from sites that have been around for at least 6 months, and that have at least 5 members, one of whom will serve as the informant.

Measures and Analysis

Audio and users' interactions with the CTools site will be recorded for analysis, using TechSmith's Morae software [website]. Digital audio will also be recorded throughout the session. Think aloud protocols from the search tasks will be analyzed, and qualitative coding analysis techniques will be used on the interview answers. Quantitative analyses will consist of counts of search task successes and failures, as well as distance measures of participants' initial guesses as to where the target documents might be located in the repository. Depending on cross-site variability, it may not be valid to directly compare task completion times across project sites using statistical techniques for continuous dependent variables; however, it should still be possible to conduct categorical analysis on counts of successes and failures, etc.

Expected Outcome

A field study is necessary because there are things I can only learn from seeing a shared repository in use by people interacting with their own documents: - How the information structure is co-constructed - Feedback that exists regarding document and folder labeling and the overall information structure of the repository, both face-to-face and through the repository - Looking for real documents in a shared repository in the field; browsing behaviors, goals, etc.

3.2 Study 2A: Labeling and Organizing

A pair of experiments will be conducted to investigate the effect of common ground and intended audience on document labeling and organizing. The design of these studies is based on Fussell and Krauss (1989). In the first experiment, participants will be asked to label and organize a set of documents provided by the experimenter, using an interface that follows the desktop file-and-folder metaphor. The task in this experiment is intended to be similar to activities members of a project group might undertake when storing and organizing information in a shared repository system.

Research Question

How do the type of common ground (from community membership or shared past experience) and the intended audience affect document and folder label choices?

Participants

Participants will be Master's students in the School of Information (MSI) who took the course SI 501 in either Fall 2006 (MSI Class of '08) or Fall 2007 (MSI Class of '09), and non-SI graduate students. Common ground in this experiment is a subject variable. Students who took SI 501 at the same time can be said to have common ground from *shared past experience*, students from SI in different cohorts have *community membership* common ground. The information they will be asked to label and organize will be related to topics in SI 501.

Method

The experiment will be a 3 (cohort) x 5 (audience) between subjects design. The levels of *cohort* are:

- MSI students in the Class of '09 who were enrolled in SI 501 in Fall 2007
- MSI students in the Class of '08 who were enrolled in SI 501 in Fall 2006
- graduate students from outside SI

The levels of *intended audience* are:

- None specified (control)
- Self (control)
- Your SI 501 class
- Graduate students in SI
- General public

Participants will be given a set of documents (related in subject matter to material covered in SI 501) to label and organize into mutually exclusive folders; i.e., each document can exist in only one place. They will be provided with instructions identifying their intended audience; i.e., information indicating who will be using the information structures they create. Each participant will create an information structure from the documents; it is expected that this structure will take the form of a tree or hierarchy. See Figure 2 on page 24 for the breakdown of conditions and number of participants in each condition. It may be necessary to match participants according to gender and English fluency; data collected during piloting will help to determine whether matching will be required.

added common
ground
manipulation
check

	No Audience	For Self	For Your Class	For Last Year's Class	For General Public	(total)
MSI Class of 2008	control 5	control 5	SPE 10	CM 10	UNK 5	(35)
MSI Class of 2009	control 5	control 5	SPE 10	CM 10	UNK 5	(35)
Non-SI Grad Student	control 5	control 5	(no subjects) 0	(no subjects) 0	UNK 10	(20)
(total)	(15)	(15)	(20)	(20)	(20)	(90)

SPE = Shared Past Experience Common Ground
 CM = Community Membership Common Ground
 UNK = Unknown Common Ground

Fig. 2: Table depicting the experiment conditions, and number of participants in each condition, for Study 2A and 2B

After organizing the documents, participants will be asked to complete a questionnaire asking them about the choices they made during the task, and including a ‘manipulation check’ for common ground. For example, the questionnaire will be used to find out whether participants report considering the information needs or expectations of their intended audience when creating the document and folder labels, and making organization decisions. The questionnaire will also include asking participants to describe their intended audience, so that it might be possible to get an idea of the model of others’ knowledge upon which their choices were based. Finally, a social network analysis questionnaire will be administered, which will serve as a proxy for common ground; the details of this instrument have yet to be finalized.

Measures and Analysis

A software application is currently being created for use in this experiment; this will make it possible for participants to organize the information via a computer application rather than using hard copies. Detailed timing data will be recorded while participants complete the organization and labeling task. The timing data will be analyzed, and a qualitative content analysis of the document and folder labels participants create will also be conducted. In addition, a quantitative distance measure will be developed to compare information structures. Finally, quantitative questionnaire data will be analyzed, and open-ended questions will be analyzed using qualitative coding techniques.

Hypotheses

It is expected that labels people create will differ depending on instructions received and common ground. Also, it should take less time for participants to organize the information for themselves than for their 501 project group, other SI students, and non-SI graduate students. It is also expected that participants will report more detailed awareness of their potential audience when they share past experience than when they share community membership. Finally, it is expected that the labels and information structures of people who share past experience will be the most similar.

Limitations and Tradeoffs

One limitation of this experiment is that common ground is a subject variable, meaning that common ground is not a treatment condition; rather, like gender it is a participant characteristic. Potential participants are classified according to the cohort to which they belong, and these categories are the levels of the independent variable. The main drawback of this technique is that it is impossible to tell with absolute certainty whether it is indeed common ground rather than a confounding variable that also changes with the levels of the independent variable causing any observed differences between groups. This is one reason common ground is often studied using referential communication tasks in the lab. Restricting the investigation of the effects of common ground to referential communication tasks, however, limits researchers to studying only *shared immediate context* common ground. The asynchronous nature of users' interactions with shared repositories necessitates the inclusion of *shared past experience* and *community membership* common ground in this experiment.

added this section

3.3 Study 2B: Information Management Outcome

In this second experiment, participants will return and find documents in the information structures (hierarchies) created in the first experiment. This experiment is designed to investigate the impact of information structures on information seeking outcomes, when the structures have been created by others with whom different kinds of common ground are shared and that were created with different audiences in mind.

Research Question

How do the influence of common ground and intended audience on document and folder label choices affect the usefulness of the information structure for finding documents?

Participants

At the end of the first experiment, appointments will be set up with participants to return 4-6 weeks later for the second experiment. Fewer participants will be required than in the first experiment, because the second experiment will include a within-subjects independent variable.

Method

In this experiment, participants from each cohort will return to the lab and complete document finding tasks in six different information structures that were created in the first experiment. Search targets will be a subset of the documents they labeled and organized several weeks earlier. The experiment will be a 3 (cohort, between) x 6 (information structure, within) mixed design. The information structures presented to each participant (within-subjects) will come from the following categories (see Figure 2 on page 24 for the table laying out all the conditions):

- Created with no audience in mind (baseline)
- Created for oneself (baseline)
- Created for oneself by someone else (baseline)
- Shared past experience: created by someone in the same SI cohort
- Community membership: created by someone in the other SI cohort
- No common ground: created by someone outside SI

clarification of
within-subjects
factor in study

Measures and Analysis

Interactions with the system while completing the search tasks will be recorded using Morae software. Measures will include time to complete tasks, number of mouse clicks, and count of wrong paths taken to find the target document.

Hypotheses

It is expected that participants will perform the best when using their own organization hierarchies, and worst when using hierarchies created by someone else with whom they share no common ground.

Limitations and Tradeoffs

The same subject variable limitation in the previous experiment exists in this experiment as well. Also, the second experiment depends upon the return of participants

added this
section

from the previous experiment. I expect a good deal of attrition, and have tried to design the second experiment using repeated measures so fewer participants will be required. Also, counterbalancing will be tricky in this experiment, as the order in which each participant experiences the information structures will have to be counterbalanced, as will the sets of documents participants try to find. Care must also be taken to vary which set of documents (search targets) is associated with which category of information structures. These details have not yet been completely worked out — it may be necessary to add 'document set' as another independent variable. Alternatively, it may be possible to create six document sets that have similar enough characteristics that this will not be necessary. These decisions will be made after the set of documents to be organized in the first experiment has been assembled, and pilot data from the first experiment has been collected.

3.4 Significance and Impact

The research described in this document will add to our understanding of language use in a situation not traditionally thought of as communication: interaction among project group members mediated by a group information management system. It will explore factors that affect the usage and usefulness of a growing category of systems to support group work. In building on previous work in both psychology and information science, I am able to investigate unstructured organization schemes from a new perspective. Results of this work will be used to suggest technological and social interventions that can be used to inform future system architecture and interface design, and development of future group information management systems.

4 Timeline and Proposed Budget

4.1 Timeline

November – December 2007: IRB approval, pilot testing of procedures finished

January 2008: Participant recruiting

February – April 2008: Data collection

Summer 2008: Data analysis

September – December 2008: Writing

January – February 2009: Finished!

4.2 Budget

Expenses for this research consist primarily of subject payments. I plan to use Morae to capture detailed task timings in the experiments, but will need to either purchase a

license or secure access to computing equipment owned by SI in order to use software licensed or purchased by the university. Subject payments will cost approximately \$3000. I am planning to apply for a Rackham Discretionary Funds grant to support this work.

Study 1: CTools Project Sites \$1100. 8-10 different sites, 1 informant (\$25) and 4 users per site (\$15). Upper bound approx \$850, plus one pilot site

Study 2A: Labeling and Organizing \$1200. 100 participants (\$10 ea.), plus several pilot sessions

Study 2B: Information Management Outcome \$700. 60 participants @ \$10 ea., plus several pilot sessions

References

- Ackerman, M. S. (2000). The intellectual challenge of cscw: The gap between social requirements and technical feasibility. *Human-Computer Interaction*, 15(2):181 – 203.
- Barreau, D. (1995). Context as a factor in personal information management systems. *Journal of the American Society for Information Science*, 46(5):327–339.
- Bates, M. J. (1998). Indexing and access for digital libraries and the internet: Human, database, and domain factors. *Journal of the American Society for Information Science*, 49(13):1185–1205.
- Bellotti, V., Ducheneaut, N., Howard, M., Smith, I., and Grinter, R. (2005). Quality vs. quantity: Email-centric task-management and its relationship with overload. *Human-Computer Interaction*, 20(1-2):89–138.
- Berlin, L. M., Jeffries, R., O'Day, V. L., Paepcke, A., and Wharton, C. (1993). Where did you put it? issues in the design and use of a group memory. In *SIGCHI conference on Human factors in computing systems*, pages 23–30, Amsterdam, The Netherlands. ACM Press.
- Blair, D. C. (2002). Information retrieval and the philosophy of language. In Cronin, B., editor, *Annual Review of Information Science and Technology*, volume 37, pages 3–50. The American Society for Information Science and Technology, Medford, NJ.
- Boardman, R. and Sasse, M. A. (2004). "stuff goes into the computer and doesn't come out": A cross-tool study of personal information management. In *SIGCHI Conference on Human Factors in Computing Systems*, pages 583–590, Vienna, Austria.
- Brennan, S. E. (1998). The grounding problem in conversations with and through computers. In Fussell, S. R. and Kreuz, R. J., editors, *Social and Cognitive Approaches to Interpersonal Communication*, pages 201–225. Lawrence Erlbaum, Mahway, NJ.

- Brennan, S. E. and Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 22(6):1482–1493.
- Bruce, H., Jones, W., and Dumais, S. (2004). Information behaviour that keeps found things found. *Information Research*, 10(1).
- Carroll, J. M. (1980). The role of context in creating names. *Discourse Processes*, 3(1):1–24.
- Chalmers, M. (2003). Informatics, architecture and language. In Hook, K., Benyon, D., and Munro, A. J., editors, *Designing Information Spaces: The Social Navigation Approach*, pages 315–342. Springer, London.
- Chang, S.-J. and Rice, R. (1993). Browsing: A multidimensional framework. *Annual Review of Information Science and Technology*, 28:231–271.
- Chantraine, Y. and Hupet, M. (1994). Efficiency of the addressee’s contribution to the establishment of references: Comparing monologues with dialogues. *Cahier de psychologie cognitive (Current Psychology of Cognition)*, 13(6):777–796.
- Clark, H. H. (1996). Common ground. In *Using Language*. Cambridge University Press, Cambridge.
- Clark, H. H. and Brennan, S. E. (1991). Grounding in communication. In Resnick, L. B., Levine, J. M., and Teasley, S. D., editors, *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association, Washington DC.
- Cutrell, E., Robbins, D., Dumais, S., and Sarin, R. (2006). Fast, flexible filtering with phlat. In *CHI ’06*, pages 261–270, Montreal, Quebec, Canada. ACM Press.
- Dourish, P. (2004). What we talk about when we talk about context. *Personal and Ubiquitous Computing*, 8(1):19–30.
- Elsweiler, D., Ruthven, I., and Jones, C. (2007). Towards memory supporting personal information management tools. *Journal of the American Society for Information Science and Technology*, 58(7):924 – 946.
- Franzke, M. (1995). Turning research into practice: characteristics of display-based interaction. In *CHI ’95: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 421–428.
- Furnas, G., Landauer, T., Gomez, L., and Dumais, S. (1983). Statistical semantics: Analysis of the potential performance of key-word information systems. *The Bell System Technical Journal*, 62(6):1753–1806.
- Fussell, S. R. and Krauss, R. M. (1989). The effects of intended audience on message production and comprehension: Reference in a common ground framework. *Journal of Experimental Social Psychology*, 25(3):203–219.
- Gordon, M. D. (1997). It’s 10 a.m. do you know where your documents are? the nature

- and scope of information retrieval problems in business. *Information Processing & Management*, 33(1):107–122.
- Greenberg, J., Crystal, A., Robertson, W. D., and Leadem, E. (2003). Iterative design of metadata creation tools for resource authors. In *2003 Dublin Core Conference: Supporting Communities of Discourse and Practice—Metadata Research and Applications*, Seattle, Washington.
- Hertzum, M. (1999). Six roles of documents in professionals' work. In *Sixth European Conference on Computer-Supported Cooperative Work*, Copenhagen, Denmark.
- Hertzum, M. and Pejtersen, A. M. (2000). The information-seeking practices of engineers: searching for documents as well as for people. *Information Processing & Management*, 36(1):761–778.
- Jian, G. and Jeffres, L. (2006). Understanding employees' willingness to contribute to shared electronic databases: A three dimensional framework. *Communication Research*, 33(4):242–261.
- Jones, W., Phuwartnurak, A. J., Gill, R., and Bruce, H. (2005). Don't take my folders away! organizing personal information to get things done. In *SIGCHI Conference on Human factors in computing systems*, pages 1505–1508, Portland, OR, USA. ACM Press.
- Krauss, R. and Fussel, S. (1991). Perspective-taking in communication: representations of others' knowledge in reference. *Social Cognition*, 9(2-24).
- Lansdale, M. (1988). The psychology of personal information management. *Applied Ergonomics*, 19(1):55–66.
- Lee, B. P. H. (2001). Mutual knowledge, background knowledge and shared beliefs: Their roles in establishing common ground. *Journal of Pragmatics*, 33(1):21–44.
- Malone, T. W. (1983). How do people organize their desks? implications for the design of office information systems. *ACM Transactions on Information Systems (TOIS)*, 1(1):99 – 112.
- Mark, G. and Prinz, W. (1997). What happened to our document in the shared workspace? the need for groupware conventions. In *IFIP TC13 International Conference on Human-Computer Interaction*, pages 413–420.
- Markus, L. M. (2001). Toward a theory of knowledge reuse: Types of knowledge reuse situations and factors in reuse success. *Journal of Management Information Systems*, 18(1):57 – 93.
- Marlow, C., Naaman, M., boyd, d., and Davis, M. (2006). Position paper, tagging, taxonomy, flickr, article, toread. In *WWW 2006 Collaborative Web Tagging Workshop*, Edinburgh, Scotland.
- Mehlenbacher, B., Duffy, T. M., and Palmer, J. (1989). Finding information on a

- menu: Linking menu organization to the user's goals. *Human-Computer Interaction*, 4(3):231–251.
- Metzing, C. and Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, 49(2):201–213.
- Mobrand, K. A. and Spyridakis, J. H. (2007). Explicitness of local navigational links: comprehension, perceptions of use, and browsing behavior. *Journal of Information Science*, 33(1):41–61.
- Nickerson, R. S. (1999). How we know – and sometimes misjudge – what others know: imputing one's own knowledge to others. *Psychological Bulletin*, 125(6):737–759.
- Pirolli, P. (2005). Rational analyses of information foraging on the web. *Cognitive Science*, 29(3):343–373.
- Polson, P. G. and Lewis, C. H. (1990). Theory-based design for easily learned interfaces. *Human-Computer Interaction*, 5(2&3):191–220.
- Rafferty, P. (2001). The representation of knowledge in library classification schemes. *Knowledge Organization*, 28(4):180–191.
- Russell, A. W. and Schober, M. F. (1999). How beliefs about a partner's goals affect referring in goal-discrepant conversations. *Discourse Processes*, 27(1):1–33.
- Schober, M. F. and Brennan, S. E. (2003). Processes of interactive spoken discourse: The role of the partner. In Graesser, A. C., Gernsbacher, M. A., and Goldman, S. R., editors, *Handbook of discourse processes*, pages 123–164. Hillsdale, NJ, Lawrence Erlbaum.
- Schober, M. F. and Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, 21(2):211–232.
- Suchman, L. (1994). Do categories have politics? the language/action perspective reconsidered. *Computer Supported Cooperative Work*, 2(3):177–190.
- Teevan, J., Alvarado, C., Ackerman, M. S., and Karger, D. R. (2004). The perfect search engine is not enough: a study of orienteering behavior in directed search. In *SIGCHI conference on Human factors in computing systems*, pages 415–422, Vienna, Austria. ACM Press.
- Trigg, R. H., Blomberg, J., and Suchman, L. (1999). Moving document collections online: The evolution of a shared repository. In *Sixth European Conference on Computer-Supported Cooperative Work*, Copenhagen, Denmark.
- Šauperl, A. (2004). Catalogers' common ground and shared knowledge. *Journal of the American Society for Information Science and Technology*, 55(1):55–63.
- Vaughan, M. W. and Dillon, A. (2006). Why structure and genre matter for users of digital information: A longitudinal experiment with readers of a web-based newspaper. *International Journal of Human-Computer Studies*, 64(6):502–526.

- Wharton, C., Rieman, J., Lewis, C., and Polson, P. (1994). The cognitive walkthrough method: A practitioner's guide. In Nielsen, J. and Mack, R. L., editors, *Usability Inspection Methods*, pages 105–140. John Wiley.
- Whittaker, S. and Hirschberg, J. (2001). The character, value, and management of personal paper archives. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 8(2):150 – 170.
- Whittaker, S. and Sidner, C. (1996). Email overload: exploring personal information management of email. In *CHI '96: Human factors in computing systems*, pages 276–283, Vancouver, British Columbia.
- Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, 46(3):441–517.