Just Email it to me!

Why Things Get Lost in Shared File Repositories

Emilee Rader

Field Prelim Paper

School of Information

August 21, 2006

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

ABSTRACT

Shared file repositories are online storage spaces where members of a workgroup can store and share files. The collaborative nature of these repositories makes them different from both one's personal information, and a large corpus like the web or a library catalog. This paper divides the various interactions that can take place with a shared file repository into four categories:

- Storing: factors affecting users' decisions to contribute files to the repository
- Organizing: factors affecting the structure (mainly filenames, tags, and locations) of the repository
- Seeking: factors affecting users' choices regarding whether to look for the information they need in the repository
- Finding: factors affecting search outcomes

For each category, factors influencing user behavior and task outcomes are presented and described, and open research questions are highlighted. Finally, several research designs are presented.

INTRODUCTION

A shared file repository is an online storage space used by a group of people who work together for organizing and sharing files. Up to 25% of a knowledge worker's time can be spent distributing, filing, and retrieving documents (Gordon, 1997). Many different kinds of workgroups including research labs, corporate teams, and software developers use shared file repositories. These repositories are maintained by many organizations, "for their potential value in the day-to-day operations of the organization" (Trigg, Blomberg, & Suchman, 1999). They are essential for document sharing (Hertzum, 1999), and can be greatly beneficial for organizational efficiency, communicating organizational goals, and also for learning and innovation (Constant, Kiesler, & Sproull, 1994). They can contain "mission critical information" such that if it were lost there would be serious consequences (Blair & Kimbrough, 2002). Despite the importance of the information stored within them, shared file repositories generally do not have explicit rules or structures for organization and searching, like a library catalog does (Blair, 2002). Instead, they tend to accumulate documents over time and become increasingly disorganized (Whittaker & Hirschberg, 2001).

A user of a shared file repository is generally familiar with her fellow group members, and with projects and joint work activities they are working on together. However, she can expect to be familiar with only some of the files stored in a shared file repository, and she may or may not have been involved with creating the hierarchy and naming structure, or with storing and moving files around in the repository. This creates a situation different from both searching the web and one's personal information, where a user might be trying to find files with which she is unfamiliar, or looking for familiar files stored in unfamiliar places. Compounding this problem is the fact that people tend to create labels for files and folders that are not descriptive or unique enough to be helpful later, when returning to the repository to find something (Blair, 2002). People also have a tendency to label and categorize files using information that is salient about them at the time they are stored, which is often not the same information that comes to mind later when we try to find the files again (Lansdale, 1988). For these reasons and others to be described in this paper, locating the information one needs in a shared file repository is a problem experienced by knowledge workers from clerical staff to senior management, who can spend significant amounts of time searching for a single missing document. Some talk about the fantasy of the "magic, psychic archive" where it is possible to always find what one needs, as quickly as possible (Kaye et al., 2006). 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).

Users of shared file repositories can be categorized by the role they play with respect to the information in the repository. Some users are "producers," file authors who create content and contribute it to the repository. Others are "consumers" who are primarily re-users of the information they retrieve from the repository. A third category of users are "intermediaries" who act as librarian or manager for the shared file repository, collecting and organizing information, and "packaging" it for others' use (Markus, 2001). In a personal repository like a laptop hard drive, the producer, consumer, and intermediary are all the same person. However, in a situation where a group is using a shared file repository this is not necessarily true. The three roles can be filled by any combination of group members, introducing problems of shared knowledge and common ground that will be discussed in this paper.

What, then, does it mean for a file to be "shared" by producers, consumers and intermediaries in an online file repository? The Merriam-Webster online dictionary (`http://www.m-w.com/`) defines the verb "to share" as "to have in common" and "to tell to others". There are of course many ways for people to share files. Often when people talk about "sharing" a file electronically they mean in the "tell to others" sense, sending and receiving via email, which is the most prevalent method used in organizations (Voida, Edwards, Newman, Grinter, & Ducheneaut, 2006). Voida et al. identified three main breakdowns users experience when exchanging information by email in an organization:

1. Forgetting which files had been shared and with whom (as a sender)

2. Finding a method of sharing that was available to everybody, with all of the desired features (for example, version control)

3. Staying up-to-date on the latest version (as a recipient)

But email is not the only way to share files. Shared file repositories are another, as are CD's and DVD's, instant messaging, hard copies, and websites, to name a few. These various methods can be classified according to three dimensions:

- push-pull: push is explicit sending from one person to another, initiated by the producer, vs. pulling information from a website or library, initiated by the consumer

- physical-digital: sharing "physical" files by handing over printed copies or CD's, vs. emailing documents or viewing web pages

- controlled-open: controlled means limiting access to specific people, for example by using permissions and some authentication method, vs. open access where anyone can view or retrieve the information

Figure 1 (next page) illustrates the relationships among several different methods used to share files in different forms (i.e. a book might be considered a hard copy of a file). For example, email and instant messaging are two methods for sharing files, that correspond to the *push*, *controlled*, and *digital* dimensions in the figure. Via email, digital files are pushed by the sender, or producer, to a specific recipient. In contrast, books and periodicals in a public library correspond to the *pull*, *uncontrolled*, and *physical* dimensions. They are retrieved by the consumer, freely available, and in hard copy form.

For the purpose of this paper, I would like to suggest that sharing files can be accomplished by making those files accessible to a specified group of people at all times, as with a shared repository (*pull*, *controlled*, *digital* in Figure 1). Sharing in this way involves separate actions by the producer and consumer, potentially with no intended recipient or use specified. This is a

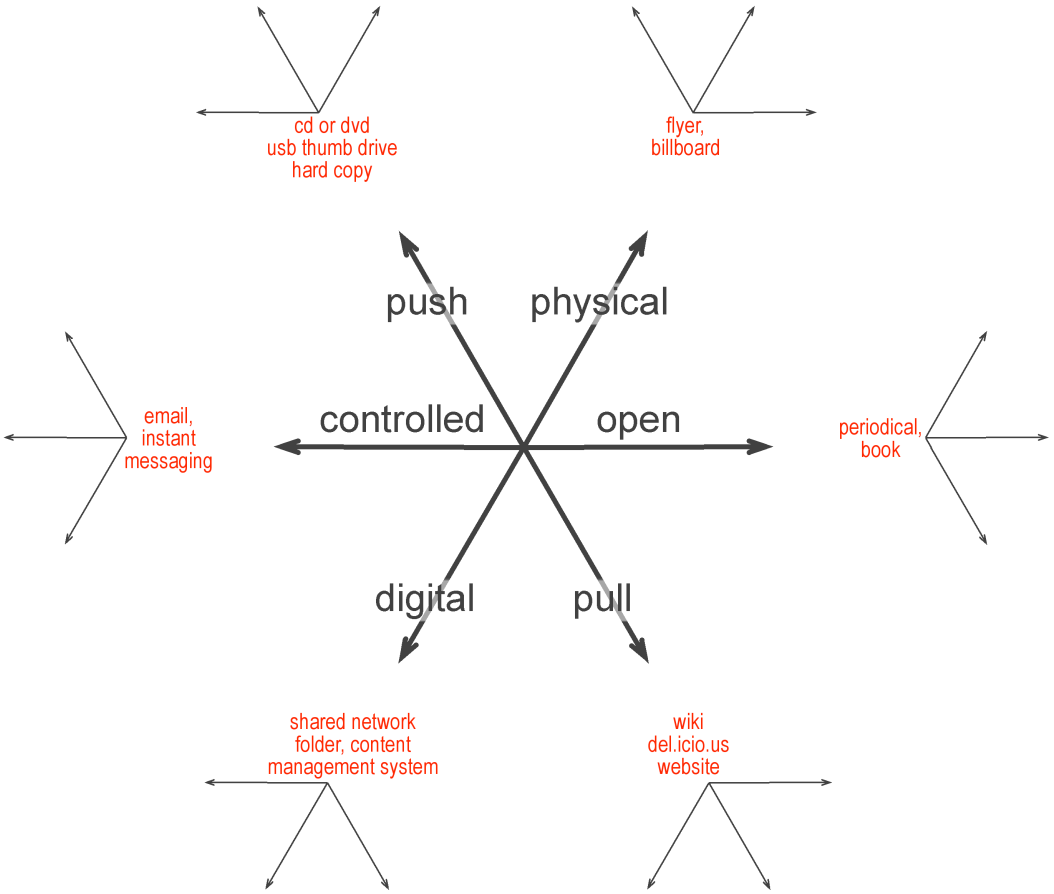

Figure 1: Classification of file sharing methods on three dimensions

slight change to the typical conceptualization of what it means to "share" a file electronically – availability and access as opposed to transmission and possession. However, this still may be considered sharing in the "have in common" sense.

Libraries and archives are two other places, in addition to shared file repositories, where groups of people are able to access shared information. However, libraries and archives differ from shared file repositories in several important respects. Libraries allow essentially unrestricted access to published works and online databases, whereas archives contain unique records that do not exist anywhere else. Shared file repositories also can contain unique information, but access is generally restricted in some way to the group or organization that owns the information. In addition, repositories tend to grow organically without an overarching plan; however, libraries and archives each employ professionals to manage the organization and growth of the collection. These differences are described in more detail in Table 1 (next page).

Shared file repositories are unique from other information repositories like libraries and archives, and also from other methods for sharing files, and therefore the solutions developed for those other areas may not be directly applicable to shared file repositories.

TYPE	PURPOSE	CHARACTERISTICS	INTERMEDIARY	EXAMPLES
Archive	permanent or long-term preservation and access	• records can consist of any media, and are unique to the particular archive (not kept anywhere else). for example: photographs, forms, memos, letters, etc. • usually contain unpublished information/records created/accumulated over the life of a person, organization, government. • content is fairly static; growth is planned and carried out by the archivist	• arranging and describing records • intermediary also provides reference services.	Presidential libraries, company intranets
Library	unrestricted access to information from many sources	• records are published works that are not unique to the institution (books, periodicals, etc) • can also provide access to online databases and other digital materials. • acquisition of new content occurs regularly; growth is planned and carried out by librarians	• develops procedures for organizing information • assists patrons with retrieval and access	Library of Congress, University Libraries, Internet Public Library
Repository	storage and (potentially restricted) access	• records/items are mostly unpublished and unique. some may be personal copies of works published elsewhere. • items are stored when they are related to personal or work-related tasks or goals for a single person or group of people • content is dynamic; growth is organic and unplanned based on individual users' actions and demands of the work	PERSONAL: the producer has full control over the contents and organization SHARED: typically no explicitly assigned intermediary	PERSONAL: active, working documents/files on one's PC; digital backups; endnote database; email; books in office SHARED: content management systems; shared network folders, restricted access websites

Table 1: A comparison of libraries, archives, and repositories for sharing information

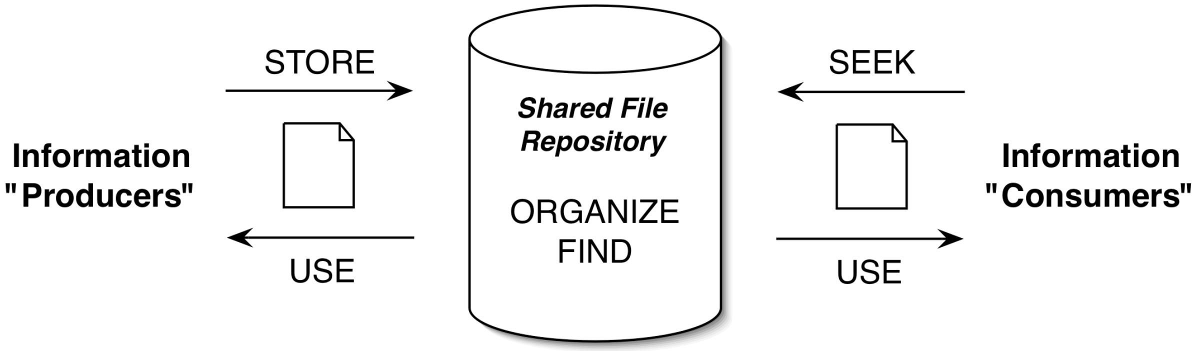

Figure 2: User interactions with shared file repositories, adapted from Markus, 2001

Users of shared file repositories have disparate roles with respect to the repository (producer vs. consumer) that often result in organization and search problems that are not experienced by individuals using personal repositories on their own computers, or conducting a library or web search. This paper will focus on the ways that users with different roles interact with a shared file repository, as illustrated in simplified form by Figure 2 (above). The problem of finding files in a shared file repository is caused by factors at many levels of analysis: individual, group, and organization. This paper will focus heavily on Organizing and Finding (center, above) at the group level, because they are the most directly related to the problem central to this paper: problems that users experience with finding necessary and important information in a shared file repository.

However, it is important to note that this is just one aspect of understanding what influences people's decisions about when and why to use (or not use) a shared file repository. There are certainly many other important areas of inquiry to be explored that are only briefly mentioned in this paper in the Storing and Seeking sections. File producers make decisions about whether or not to store their files in the shared file repository; these decisions are affected by factors such as their willingness to share what they have produced, the norms and incentives of their group regarding the shared file repository, and the degree of interdependence inherent in the work. File consumers, on the other hand, must decide whether to look in the repository for the information they need or try to find it by some other means – hence the title of this paper, "Just email it to me!" Incentives, critical mass and network externalities affect that decision, as well as the relevance of the information in the repository to the information need, and the ease of use of the repository.

In the sections that follow, I will discuss separately the four major interactions that users have with shared file repositories (store, organize, seek, find) in detail, identifying what is currently known and suggesting avenues for future research. To this end, in each section I will call attention to research questions as they arise, using the convention below:

Q0.1: Why is the sky blue?

In the "Research Questions and Designs" section, I will select a small number of these research questions, discuss the reasons behind my selection, and propose several research designs to address those questions.

CONCEPT MAP: Organizing and Finding

The majority of this paper focuses on the concepts and relationships represented by the diagram on the next page. This diagram assumes several things:

- A workgroup that is using a shared file repository
- Information producers that have decided to store files in the repository; for a discussion of factors that affect this choice, see Part I: Storing (page 12)
- Information consumers who already have a pretty good idea about their information needs, and have decided to search the repository; more information about factors affecting this choice can be found in Part III: Seeking (page 24).

The left side of the dotted line in the diagram represents factors that contribute to the ways in which a shared file repository is *organized*, the external representation, which is assembled over time as information producers choose filenames, labels, tags, and positions in a folder hierarchy for the files they add. The external representation includes both the functionality of the application software, and the underlying organization or structure of the information contained within the repository. "Location" in the diagram does not only refer to an absolute path to a file in a system that does not support multiple classification; a "location" can also be metadata assigned to a file, intended as the way in which a future information consumer might gain access to that file. Concepts in Part II: Organizing (page 15) include theory from the psychology of language and categorization, and findings from personal information management and information behavior research.

The external representation, in the center of the diagram, is a bridge between the information producers and consumers. It is an artifact shaped by the processes and factors included on the left side of the diagram, and that influences the internal representations of information consumers, shown on the right side of the diagram. On the right side of the dotted line are factors that contribute to whether or not the information a consumer is searching for will be *found*. In Part IV: Finding (page 28) factors including the structure of a consumer's internal mental representation of the information in the hierarchy, and the possible effects of different external representations on search outcomes will be discussed.

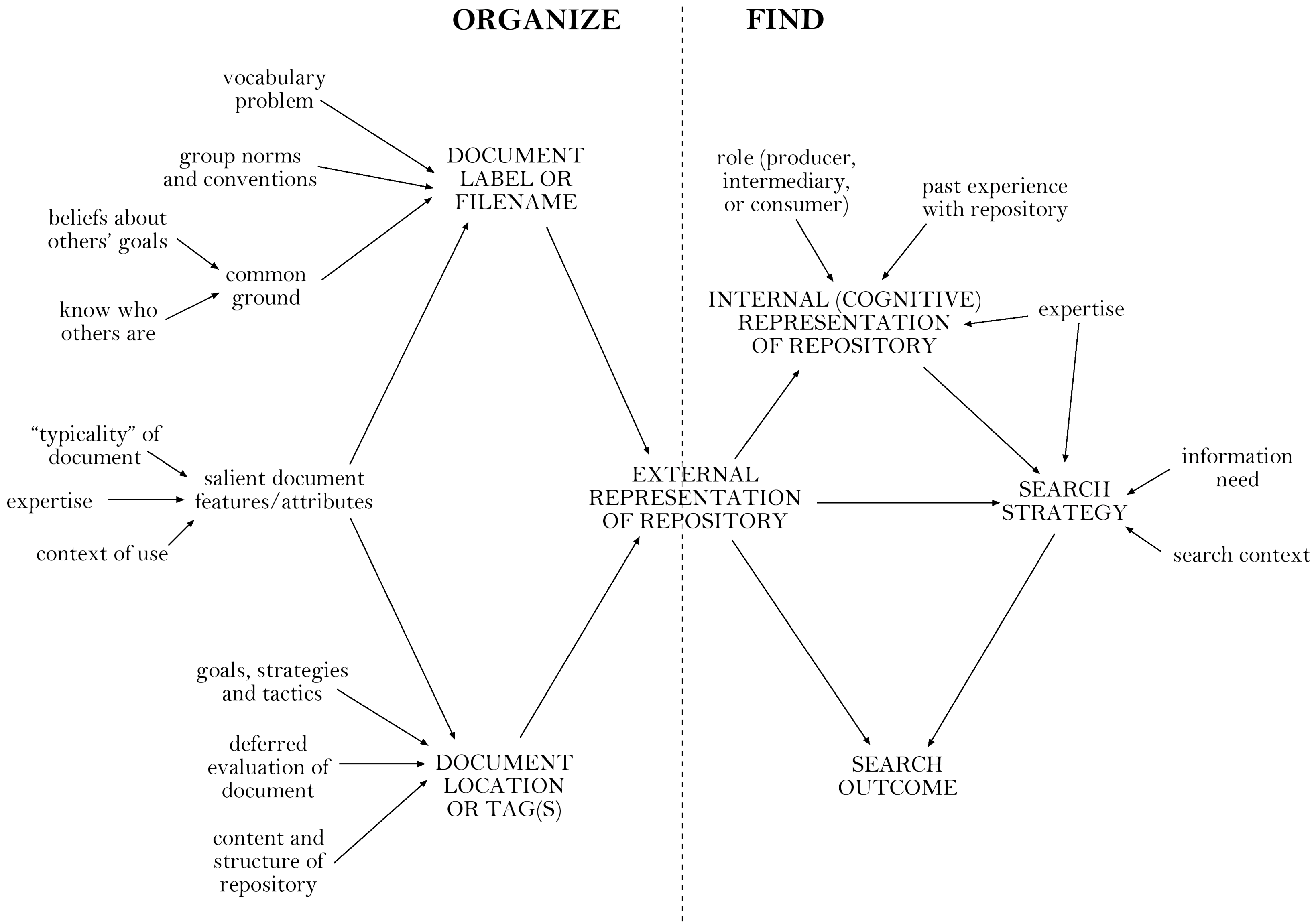
ORGANIZE
FIND
vocabulary problem
group norms and conventions
beliefs about others' goals
common ground
know who others are
DOCUMENT LABEL OR FILENAME
"typicality" of document
expertise
context of use
salient document features/attributes
goals, strategies and tactics
deferred evaluation of document
content and structure of repository
DOCUMENT LOCATION OR TAG(S)
EXTERNAL REPRESENTATION OF REPOSITORY
role (producer, intermediary, or consumer)
past experience with repository
INTERNAL (COGNITIVE) REPRESENTATION OF REPOSITORY
expertise
information need
SEARCH STRATEGY
search context
SEARCH OUTCOME

Figure 3: Concept map depicting possible relationships among factors affecting shared file repository organization and search

PART I: Storing

In groups that use a shared file repository, information producers must decide whether or not they will participate and store their files in the repository. There might be many motivations for this decision: to archive or backup the information so it will be preserved (Kaye et al., 2006), to make the information accessible for consumers (or themselves) at a later time, or perhaps to conform with group norms for information storage and sharing. This section will introduce some individual- and group-level factors that might contribute to these decisions, and suggest some areas for further research. Subsequent sections will address other interactions with shared file repositories (see Figure 2, page 9).

One factor that might influence an information producer's use of a shared file repository is the type of work being done, which could affect whether or not he is motivated to make his work products available to others via a shared file repository. In a social psychology literature review, Wageman (2001) focused on one characteristic, interdependence, that is found at different levels in different types of work. She described two types of interdependence, task and outcome. *Task* interdependence has to do with features of the work itself that require it to be done by multiple people. For example, a large software development project worked on by multiple programmers whose code must interoperate exhibits task interdependence. *Outcome* interdependence occurs when shared incentives or rewards exist that depend on the performance of the group. An example of outcome interdependence might be an annual bonus that is calculated based on how well an entire sales team has met their goals for the year. According to Wageman, task interdependence more so than outcome interdependence directly influences cooperative behavior in teams. Thinking about the examples of a programming team (task) and a sales group (outcome), the programmers are in a situation where their code must work together, or the software development project is a failure. In contrast, as long as the salesmen continue to make individual sales they are performing adequately, and in fact the annual bonus may encourage each of them to work harder individually. They don't need to cooperate to get a higher bonus.

Wageman's results suggest that groups like the programming team that exhibit task interdependence might be more likely to contribute to a shared file repository than groups like the sales team that exhibit outcome interdependence. Teams with task interdependence have more reason to collaborate on work products, and therefore could be more likely to share files. However, interdependence is only part of the picture. Whether or not information producers choose to store files on the shared file repository might also depend on how willing they are to share information with others.

Constant, Kiesler and Sproull (1994) asked what affects people's willingness to share information with others in their organization. They found that if the norms of the organization are such that workers believe the *organization* rather than the *individual* owns the work products, people are more likely to be willing to share them. Information work products that might be owned by the organization or the individual, depending on norms, are things like, "an idea, process, invention, document, or computer program" (p. 404) created at work or using organizational resources (Constant et al., 1994). For example, papers written by an employee of a large corporation might be perceived as being owned by the corporation, while papers written by a graduate student employed by a university are more likely to be thought of as belonging to the student. If Constant et al. are correct, then the employee of the corporation might be more willing to share their work via a shared file repository than the student.

Considered together, the findings of Wageman and Constant et al. seem to indicate that groups with high task interdependence and norms for organizational ownership would be more likely to contribute to a shared file repository than groups with outcome interdependence and norms for individual ownership.

> Q1.1: What are the effects of interdependence and ownership norms on information producers' attitudes toward sharing, and contributions to a shared file repository?

It is also important to consider *incentives for participation* as a factor affecting whether information producers decide to contribute their work to a shared file repository, which is essentially a collaborative system (Ackerman, 2000; Jian & Jeffres, 2006). After all, what is in it for them – what do they get out of spending their time and effort deciding what to store online, packaging it so others might understand it (Markus, 2001), and interacting with the repository?

Social psychologists Karau & Williams (2001) developed the "Collective Effort Model," identifying performance- and reward-related factors that contribute to the level of an individual's participation in group work. The model suggests that people will contribute to a collective task only when their contributions result in an outcome that they value personally. However, it is less obvious when their own effort and contributions lead directly to these personally valued outcomes in collective work, than individual work. Said another way, it is as if people try to estimate how much of their effort has contributed to the outcome, and in collective tasks it can be hard to tell. Karau & Williams suggested that in these instances, people don't try as hard, and loafing occurs. In a shared file repository this might result in a pattern where a small number of people contribute a lot of content to the repository because they find it personally rewarding to do so, and a larger number of people contribute very little because they don't perceive any personal benefit. Markus (2001) refers to the challenge of designing appropriate incentives for repository contribution as the "discretionary database problem", explaining that organizational incentives (rewards) may be required to encourage people to participate, especially when they are pressed for time or directly competing with each other (Olson & Olson, 1999).

Palen & Grudin (2002) studied the adoption of online calendaring applications in the mid-1990's, an example of an incentive problem that has a lot in common with shared file repositories. Their research was a follow-up to an earlier classic paper by Grudin (1988), describing the discrepancy in incentives that existed in online calendar use. In the earlier paper, Grudin wrote that users who benefited the most from the data that was entered into online calendars were people like managers who had secretaries to do a lot of the data entry and meeting scheduling work for them. The lower-level employees did not enjoy such personal benefits, and therefore were reluctant to use the software. Without their data in the system, the online calendars were useful to no one. Interestingly, Palen & Grudin found that in the mid-1990's the incentive barrier to online calendaring no longer existed. The calendars were in widespread use in both high tech companies they studied. They found that this was due to several changes in the software itself, and in the norms of the organizations. The functionality in the generation of calendar software they studied was more useful and better designed, tech support was better, and everyone used the same software so all calendars could interoperate. In addition, it had become social convention for everyone to use the calendars. Once everyone was using them, it was no longer just the managers who received the benefit from having scheduling information online.

Information producers are presumably more motivated to store information in a repository that they themselves would want to access later, than purely for the benefit of others. Markus (2001) wrote:

> "Producers have the greatest natural incentives to create repositories that benefit themselves directly in use. They have some but lower natural incentives to create repositories for similar others (in the shared work practice situation), where they can potentially benefit from others' reciprocity. They have lowest natural incentives to document for dissimilar others, where the primary reward is the user's gratitude" (p. 84)

An implication of the findings of the online calendaring studies is that any personal benefit information producers receive from storing files in a shared file repository depends upon the contribution of files by others, or on the need to access their own files again (which probably exist in a personal repository somewhere anyway). Like online calendars, shared file repositories are only as useful as the information they contain – if nobody enters their schedule, there is no reason for an individual to check the calendar when putting together a meeting. Similarly, if nobody is storing useful content in a shared file repository, there is little reason for any particular individual to visit the repository at all, unless it is being used exclusively for backing up information. In both cases, valued personal outcomes result more from information that is entered into the system by others, than from one's own contribution. This presents a bit of a chicken-and-egg problem – how does use of a shared file repository ever get started in the first place? In the online calendar example, better software and ubiquitous deployment was the solution. It is not clear whether this might work for shared file repositories.

> Q1.2: What factors lead to the successful adoption of shared file repositories by information producers?

Also, shared file repositories are different from the calendaring example in one important way: there are different types of repository users: producers, consumers, and intermediaries. It is not necessary for a majority of consumers to also be producers for the system to contain enough content to provide value. It is conceivable that many information consumers would find personal benefit from using a shared file repository where only a few producers are storing files. This seems to be the case in many Usenet listservs, where a few prolific posters support a huge number of lurkers.

> Q1.3: What is an optimal ratio of information producers to consumers to sustain continued use of a shared file repository? How is this affected by the purpose for which the repository is used, and the type of content stored?

Finally, a very important individual-level factor contributing to whether or not an information producer will store files on a repository is the level of trust she has in the technology (Berlin, Jeffries, O'Day, Paepcke, & Wharton, 1993). Technical difficulties and poor usability do not inspire confidence that the information will still be accessible when someone else needs it again. Whittaker & Hirschberg (2001) believe lack of trust is a significant obstacle to the use of shared file repositories, saying, "mistrust of public stores means that a global repository managed by others would not be acceptable" (p166).

In summary, several factors were introduced in this section that may affect whether information producers choose to store files in a shared file repository (see Figure 4, below). In a group with norms for organizational ownership and high task interdependence, one would expect to see high participation in a shared file repository. However, incentives and trust still play a role; there must be enough perceived personal benefit, and the users must trust the system enough to store their work there. The following sections will address the decisions made by information producers regarding the location and labeling of files in a shared file repository, and the information behavior of information consumers with respect to searching for files in a repository.

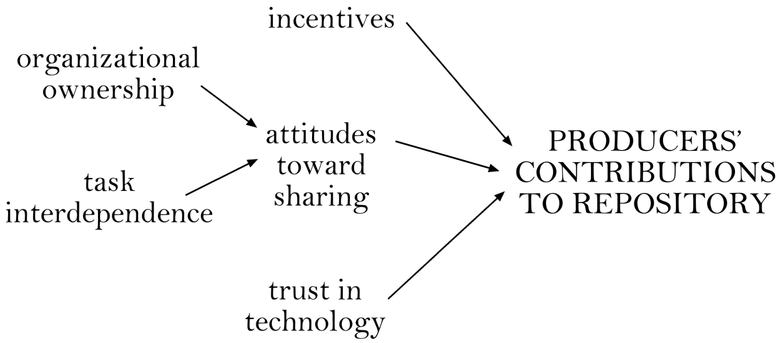

Figure 4: Factors influencing producers' contributions to repository

PART II: Organizing

This section explains some of the factors contributing to the selection of labels and filenames, and decisions information producers make about where to store files in a shared file repository (Figure 3.1, next page). When an information producer adds a file to a shared file repository, she gives that file a name and either puts it in a specific location within a folder hierarchy, or assigns to it some metadata. These are two nearly universal decision points encountered in the course of adding content to a repository. There is a great deal of variability among the choices two information producers might make (Furnas, Landauer, Gomez, & Dumais, 1983; Markus, 2001); these decisions shape the structure of the repository, and impact information consumers' ability to find the files they need (Mark & Prinz, 1997). Maintaining a shared file repository is a collaborative activity, and the repository is more complex than just an "aggregate of every individual's contribution" (Jian & Jeffres, 2006). Hertzum (1999) wrote, "Document management must strike a balance where it inflicts minimal inconvenience on the individual professional, and yet ensures the quality of the shared archive" (p56). This section addresses how some of that inconvenience arises.

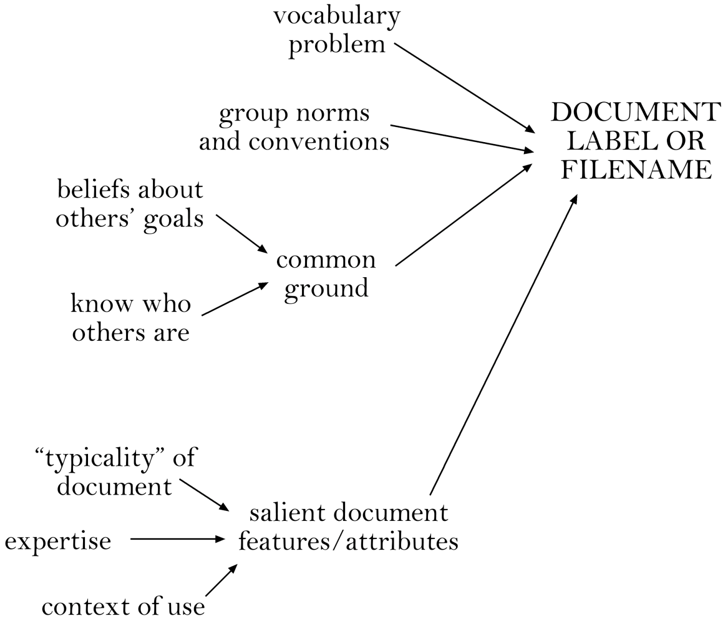
vocabulary problem
group norms and conventions
DOCUMENT LABEL OR FILENAME
beliefs about others' goals
common ground
know who others are
"typicality" of document
salient document features/attributes
expertise
context of use

Figure 3.1: Factors affecting labels or filenames

Document Label or Filename

Conventions

Markus (2001) related an example about a team using a shared file repository system, who encountered one common problem with conventions related to repository use:

> "The team used a collaboration technology with sophisticated keyword search and retrieval facilities. At the outset of the project, team members jointly agreed to document everything that might be of use to them later and to assign at least three keywords to each document in the database. The team immediately became overwhelmed by the demands of documentation and soon settled for less documentation and more synchronous interaction" (p80).

In this example, members of the group using the repository decided upon conventions for keywords before they began using the system, and later found out that what they had decided upon was inappropriate for the situation at hand. *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 & Fussell, 1991). In another example, 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 & Prinz (1997) conducted a field study of a group using a "large groupware system" to store and share files. 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 & Prinz 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 (Mark, 1997; Mark & Prinz, 1997). One user said, "Naming conventions, reference code, and subject area, I always violate. I give file names that seem to fit" (Mark, 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; Berlin et al., 1993; Mark & Prinz, 1997). Shared file 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 files 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. If the normal mechanisms for developing and enforcing social conventions do not apply to shared file repository systems, then:

> Q2.1: How do file naming and labeling conventions evolve in shared file repositories? How are they enforced, or reinforced?

Vocabulary Problem and Common Ground

Furnas et al. (1983) described what they would come to call the *vocabulary problem.* They reported that random pairs of people use the same label for an object at most 20% of the time. They wrote:

> "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 file repositories are dire: if two random users were to create a label for the same file, they would be far more likely to choose different labels than the same label. Similarly, if an information consumer attempts to imagine what the file he is looking for might be called, chances are low that he will end up looking for the correct filename. Fortunately, users of shared file repositories are not necessarily random pairs of people who are unknown to each other. In the best case, they could share a work context and even have some knowledge about each other's preferences and personal styles. Humans' use of language is imprecise and flexible, and meaning is determined by the surrounding context, and complex communication processes. While a shared file repository is not a communications system, language is being used as abbreviations to represent the contents of files (labels/filenames), and also to suggest relationships among groups of files (folder names or metadata). So, while the vocabulary problem introduces a great deal of variability into the labels that people generate, their knowledge about each other and their shared context – their *common ground* – might mitigate the problem somewhat, if it were brought to bear.

Common Ground (Clark, 1996)is a psychology theory describing conversational processes that enable people to understand each other while they are talking to one another. Common ground is defined as the mutual knowledge, beliefs and assumptions that conversation participants share about each other. It is inferred based on joint membership in cultural communities and through shared perceptual experiences, and accumulates via conversation. An essential characteristic of common ground is that it is reflexive, meaning that it does not just consist of what conversation participants know or assume about each other; it also includes the beliefs and assumptions they each make about what the other knows about them (i.e., "I know that you know that I know…" and so forth). As 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. According to Clark, common ground is necessary for coordination of conversation, and essential for people to understand one another. For example, two people (A and B) who grew up in the same city share common ground that is not shared with a third person, C, from a different city (Clark, 1996). A and B become aware of their common ground through conversation, i.e., this common ground does not exist between them until they are both aware that the other is from the same city. Because of their common ground, when talking about the location of the closest coffee shop A and B are able to refer to landmarks that C is unlikely to be aware of.

Conversation participants believe common ground exists when there is evidence for a "shared basis". In the examples above, the shared basis is growing up in the same city, or being present in the same classroom at the same time. Evidence that a shared basis exists for members of a workgroup using a shared file 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.

> Q2.2: What evidence for a "shared basis" or "feeling of others' knowing" exists and can be communicated in a shared file repository?

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 & 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 'overhearer') 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 & 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 & 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 names information producers create for files they store in a shared file repository. People tailor what they say to whomever is the intended recipient, even when they are simply instructed to write descriptions for "someone else". Groups with more common ground might label files with names that others in the group will be able to anticipate more than 20% of the time. However, this is not as straightforward as it sounds. Hertzum & 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 file repository. The ability to take the perspective of others is also necessary.

Q2.3: How does common ground affect information producers' choices of labels for files?

Q2.4: To what extent do information producers "package" their contributions to a shared file repository? How does this affect repository use by information consumers?

Interestingly, this problem does not occur exclusively in shared file repositories. It even occurs between professional catalogers and information seekers. Šauperl (2004), interviewed 12 catalogers about their process for cataloging, and concluded that they were more concerned about common 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 (2004) 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 items have been cataloged in the past.

Salient Document Features

Many people can identify with the experience of losing an object like one's keys, and while looking for them thinking, "Now what was I doing the last time I had them?" Not only do we remember specifics about the sought-after object; we also remember other information about the surrounding context in which it was used. Context causes certain information to be *salient*, or visible and important; this affects where a file might be stored in a shared file repository, and how an information consumer might choose to search for it (Lansdale, 1988). For example, one might remember using the keys to open the door upon returning from the grocery store, but not remember moving them from their pocket to an end table while watching an exciting football game on TV. When people make decisions about where to file things in a repository, situation attributes such as surrounding circumstances, anticipated need for the file, source, etc. often determine the location. However, as projects and priorities change, salient situation attributes can also change. (Barreau, 1995; Kwasnik, 1991). Whether an information producer is labeling, tagging, or selecting a location for a file, the context of use affects his choices.

Another way that salient attributes of a file can change over time and affect where the file is stored is as the file falls into disuse. Hertzum (1999) wrote about how documents move from "action to archive", from "spatial, loosely systematised, memory-based organization to category structures" (p. 43) as they become less and less important to daily work activities. Files that are frequently used tend to be stored in places that are quickly and easily accessible; older information that is used less often is organized into more concrete and complex categories that are less visible (Barreau & Nardi, 1995; Malone, 1983).

Jacob (1995) wrote, "category membership is not contingent upon a set of shared attributes or properties but is determined, instead, by the individual's recognition of integral relationships that exist between the observable properties exhibited by a set of category members." This refers to the idea that categories – folders in a shared file repository – are created based on perceived relationships among the information in the files. However, these relationships and the perceptions of them are not stable over time, even within the same information producer. In addition, 'typical' files for which there is an obvious category are easy to assign to a category. But for atypical files that could exist in multiple different categories, people tend not to agree on where they should go because they choose different features upon which to base their category assignments (Mervis & Rosch, 1981). Finally, research has shown that categorization differences exist between experts and novices in a domain (Kellogg & Breen, 1987; Marchionini, 1997; Murphy & Lassaline, 1997). Expertise might cause differences in the perceptions of the relationships among the files in a shared file repository, and also how these relationships are represented by the categories to which they are assigned. This could make it difficult for experts and novices to use each other's file structures.

Document Location or Tags

When storing a file in a shared file repository, labeling is only half of the task. Information producers must also decide where the file will be stored, i.e., the location at which someone else will be able to access the file again. This "location" can be an absolute position within a hierarchy, or it could be user-generated metadata (tags) assigned to the file for the purpose of aiding in retrieval (discussed in more detail in Part 4: Finding, on page 32). The figure below (Figure 3.2, next page) depicts factors that affect information producers' choices.

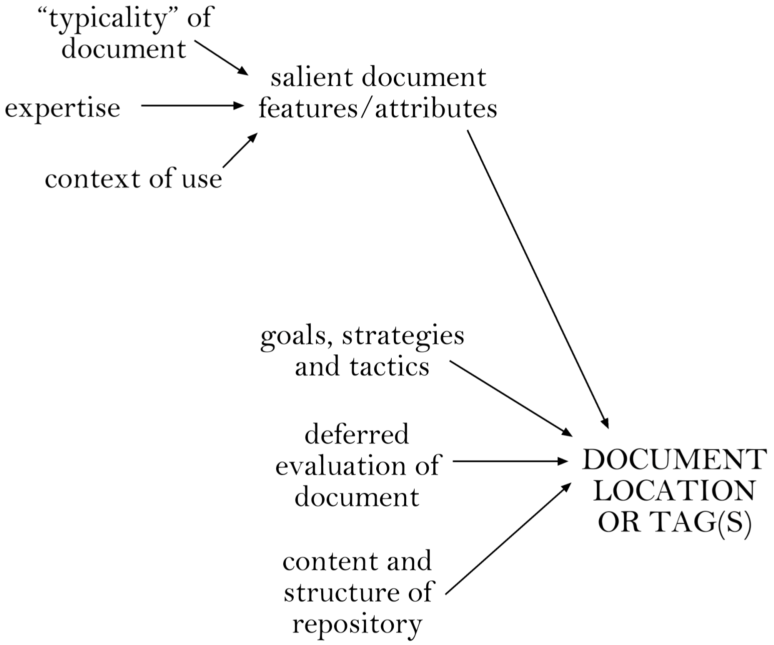
"typicality" of document
expertise
context of use
salient document features/attributes
goals, strategies and tactics
deferred evaluation of document
content and structure of repository
DOCUMENT LOCATION OR TAG(S)

Figure 3.2: Factors affecting document location or tags (excerpt from concept map)

Deferred evaluation, pruning, and types of information

Whittaker & Sidner (1996) wrote that filing is a "cognitively difficult task", because when an information producer is deciding where to put a file, he must imagine where he and others might want to go looking for the file again, as well as remembering how everything else is classified, 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 file 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, Ducheneaut, Howard, Smith, & Grinter, 2005; Malone, 1983; Whittaker & 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 items, there are a lot more folders to remember than if each one contains ten or fifteen items. Kaye et al. (2006) wrote "...keeping a clean and compact digital space meant a minimum of mental overhead to track items within it" (p282).

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 files 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 file is put aside, it probably won't happen at all (Whittaker & 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 file should be stored in the repository or not, meaning the files might not end up in the repository at all.

When trying to decide what to do with a new or incoming file, there are three high-level problems a person faces (Whittaker & Hirschberg, 2001):

- figuring out the "value" of incoming information, whether it is important or needed
- figuring out how to categorize the information
- deciding where to put the information, or deferring judgment

These findings came from a study of personal information management, but there is no reason to suspect that they would be invalid for shared file repositories. In fact, users of a shared file 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. Shared file repositories are like the refrigerator, without the mold. Old files are a lot less offensive than moldy pizza, so the motivation to purge a repository is less likely than for the workplace fridge. Also, the evaluation decisions are harder in a shared file repository. Clearly, nobody will want the moldy pizza; but the choice may not be so black-and-white for old meeting minutes or out-of-date lab procedures. The path of least resistance is to leave things as-is.

There is one other difference that comes to mind when filing or categorizing in a group setting. Suchman (1994) cautioned that categorization serves not only to make things more organized; it can also communicate information about the values of a group, and in essence be a form of social control. Document labels and the representation of the relationships between content items 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. In a response to that paper, Winograd (1994) argued that "...structure is not an imposition of control for authoritarian motives, but a necessity of continued operation..." (p. 95), meaning that as an organization grows, more structure is required to keep things running smoothly, and that this is not necessarily an evil thing. Their perspectives hint 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.

Goals and Strategies

People have different ways of structuring their personal file repositories, as has been observed in many personal information management studies (Barreau, 1995; Berlin et al., 1993; Boardman & Sasse, 2004; Kaye et al., 2006; Malone, 1983; Whittaker & Hirschberg, 2001; Whittaker & Sidner, 1996). These range from "save or purge" to "frequent filers" vs. "pilers", to "I do what works for me." The specifics mentioned in all of this previous research are not important, but the overarching finding is: when managing information, user behavior can be broken down at different levels of granularity into "goals, strategies, and tactics" (Bellotti et al., 2005; Marchionini, 1997). Thinking about behaviors in this way separates the behavior from the person, so that individual differences aren't solely responsible for all of the variability, and the researcher can learn about meaningful patterns. Instead labeling people as "filers and pilers" and leaving it at that, one can look at goals for which certain strategies are used more than others for example, test hypotheses, and draw conclusions.

- Goals: The highest level of granularity, things like "keep my email organized (Marchionini, 1997) or "making deadlines, managing information as it comes in, distributing information, being prepared for events" (Bellotti et al., 2005).

- Strategies: Choices users make for how to achieve their goals, such as "keep all to-do messages in my inbox" (Marchionini, 1997) or "keep relevant content at hand" (Bellotti et al., 2005), or "re-use tags to maintain consistentcy".

- Tactics: individual actions that are undertaken in support of strategies, like "filtering messages by sender to find a particular message with an important attachment" (Marchionini, 1997) or "look through tags I've already used to see if I can use any of them for this file". Bellotti et al. (2005) wrote, "At the tactical level, we see an enormous amount of variation that reflects user exploitation of the complexity and particular characteristics of modern email tools and other computer resources" (p102).

Groups find it difficult to keep shared file repositories organized when multiple people are acting according to their own idiosyncratic strategies. This affects an information consumers' expectations for whether a particular file is likely to exist in the repository, what that file might be called, and where it might be located (Berlin et al., 1993; Mark & Prinz, 1997).

Q2.5: What are the goals of information producers and consumers, related to their use of shared file repositories?

Q2.6: How do the strategies and tactics of information producers affect the structure of the shared file repository, and the strategies and tactics of information consumers?

There are implications of these findings from studies of personal information management and psychology for the structure of shared file repositories. The goals and strategies of individual information producers for managing information could affect the choices they make when contributing to a shared file repository. These choices are also likely to be affected by personal preferences for simple structure; a person with a high need for simple structure might be more likely to put files in folders that already exist than to create new, additional folders, for example. Or, they might be more likely to take on the role of an intermediary responsible for reorganizing the repository.

Summary

This section has discussed several factors that might affect the labels selected for files in a shared repository, and the locations chosen for them. File naming conventions, even when they are agreed upon in advance, are hard to stick to in practice. They often do not reflect the realities and constraints of the work, and individuals intentionally or unintentionally stray from them. In addition, the feedback and cues necessary to reinforce conventions are absent from the user interfaces of shared file repositories. Two random people are unlikely to choose the same label for the same file; however, common ground might improve the likelihood that this would happen – if information producers are thinking about information consumers when making labeling and filing decisions. Context and expertise can affect salience of different aspects or information about a file; this can affect both how it is labeled, and where it is stored. Also, files that are typical when compared with other information in the repository are likely to be filed more consistently than files that are atypical, which could be stored in many places. Users' goals, strategies, and individual differences can also affect where files are stored. Finally, information producers don't like to prune or purge, and therefore shared file repositories tend to accumulate content and grow larger without the benefit of a consistent plan for organization.

PART III: Seeking

The previous section described how information producers go about labeling and organizing the information in a shared file repository. But information producers are only some of the users of the repository – information consumers are also an important part of the picture (see Figure 2, page 9). Just as there are factors that affect whether an information producer will decide to use the repository, there are factors that influence the behavior of information consumers as well. This section suggests things that might affect whether a consumer looks for information in a shared file repository, and what they might be looking for (Figure 5, next page).

Willingness to Use

In order to consider searching a shared file repository, an information consumer must be aware that the repository exists, have available the requisite technology for access, and have the knowledge they need to use the repository effectively (Leckie, Pettigrew, & Sylvain, 1996). The consumer's information need also affects his decision to seek out the information in a repository, vs. some other way of obtaining the information. As part of this evaluation, the information consumer weighs the benefit he would expect to gain against his effort to use the repository. McCreadie & Rice (1999) wrote that this cost-benefit analysis is "influenced by resources available such as motivation, time, convenience, level of tolerance for uncertainty, delayed gratification or inconvenience or a world view that sees the potential for addressing the situation as likely" (p. 60).

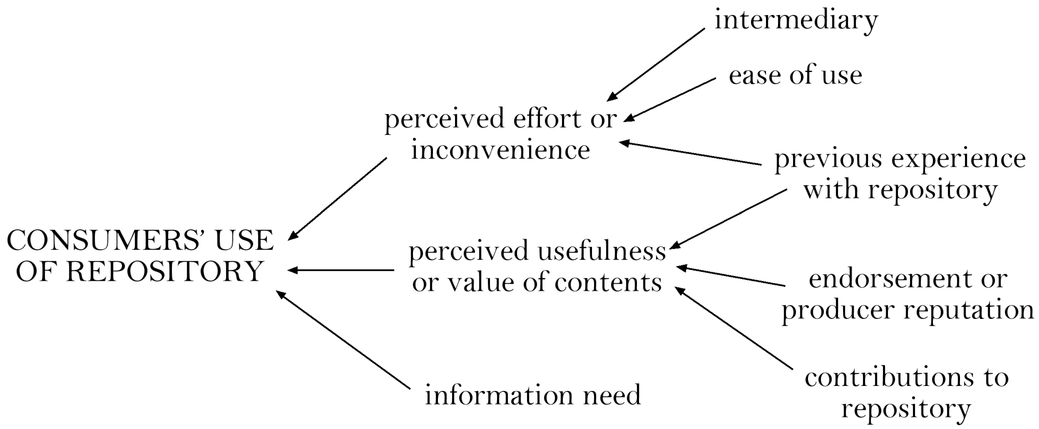
intermediary
ease of use
perceived effort or inconvenience
previous experience with repository
CONSUMERS' USE OF REPOSITORY
perceived usefulness or value of contents
endorsement or producer reputation
contributions to repository
information need

Figure 5: Factors influencing consumers' use of repository

The benefit is also directly related to the number of others contributing to the repository. Critical mass is certainly an issue, at least initially as with any collaborative system (Ackerman, 2000). There must be some number of contributions to the repository for it to be viable; this number likely depends on the purpose for which the repository is used. In addition to critical mass, network externality effects are more important. The personal value obtained from participating in a shared file repository increases as more files are contributed to it (Star & Ruhleder, 1996). For example, one can imagine that as more files are added, it is more likely an information consumer might find a useful document there. People are more likely to search a shared file repository if they believe they will find something useful (Kankanhalli, Tan, & Wei, 2005). However, there might also be diminishing returns; at some point there might be too many files that are too disorganized for an information consumer to find what they need.

When thinking about information in a shared file repository, one might wonder, where does the most useful information for a workgroup come from? According to Levin & Cross (2004), more "useful information", defined as information that has a positive impact on an information consumer's work, comes from weak ties in an organization than from strong ones. The idea behind this is that the people one works with day in and day out rarely offer some previously unknown tidbit of information that turns out to be extremely useful. However, people one interacts with less often are more likely to contribute novel information that turns out to be very useful. This is the same phenomenon behind the observation that when hunting for a job, acquaintances are more likely than friends to provide a key lead. Levin & Cross operationalized tie strength by an assessing the closeness of working relationships, and communication frequency, via a survey instrument. They found that knowledge from weak ties contributed more positively to project outcomes than did knowledge from strong ties (which also contributed positively). Gordon (1997) mentions the same phenomenon, saying, "information storage/retrieval/sharing systems can have their greatest effects across departmental boundaries within a given organization by allowing better exchange among people who would ordinarily be isolated from each other" (p119). This seems to suggest that an information consumer might find information contributed to the repository by weak ties more useful than the contributions of strong ones, and therefore might be more willing to use a repository that connects weak ties together. Perhaps a social network analysis of participants in a shared file repository to identify the number of strong vs. weak ties might be used to predict level of participation by information consumers.

> Q3.1: Does the strength of the ties in the social network of shared file repository users predict the level of participation by information consumers?

Markus (2001) suggested two other factors that could contribute to consumers' willingness to look for information in a shared file repository: personal referrals to specific documents (essentially endorsements), and familiarity with the reputation of a file's authors. This additional information about the contents of the repository presumably allows information consumers to make more accurate judgments about the quality and usefulness of the information they might find there. In fact, a study of the information behavior of engineers found that participants used the corporate archive more for finding pointers to people from whom they would then solicit the needed information, than for finding documents that contained the information (Hertzum & Pejtersen, 2000). So it seems that sometimes information about the contributors to the repository is more useful than the information contained within the repository.

An information consumer's expectations for what a repository contains are influenced by her previous experiences with the repository. Her memory for the information she has encountered there in the past influences her likelihood of searching the repository given a specific information need. For example, an information consumer who has first-hand knowledge of only a subset of the files in the repository might be likely to assume that the rest of the repository contains items similar to those that she uses frequently. Information that is recalled more easily (i.e. info that is more "available") is judged by people to be more common, frequent or well-known than information that is difficult to recall (Tversky & Kahneman, 1973). So, the information consumer might not ever think to look for data and research papers in a repository she has mainly been using to fill out forms for reserving equipment. She would probably believe that the repository contains only items similar to those she had found there in the past, or those that others had talked about using (Wright, Mathews, & Skagerberg, 2005). This bias could influence her willingness to look for other kinds of information in the repository.

Finally, the intervention of an intermediary, a third category of shared file repository user, might increase the likelihood that a consumer would look for information in a repository. Intermediaries are responsible for collecting and organizing information, and "packaging" it for others' use. Intermediaries aren't always an officially recognized position; sometimes the need for one is so great that an information producer or consumer who is a member of the team steps forward on their own to fill the role (Berlin et al., 1993; Kaye et al., 2006). Markus (2001) wrote that contents of repositories "differ depending on whether the record keepers are knowingly documenting only for themselves, for others who are similar to themselves… or for others who are dissimilar" (p. 72-73). This can result in information needed by consumers either not existing in the repository, being difficult or impossible to find, or lacking the necessary context for interpretation. The efforts of intermediaries toward maintaining an effective repository could mitigate some of these problems, and make it a viable resource for information consumers.

Information Needs

There have been no recent qualitative studies describing seeking and retrieval patterns for shared file repositories. A few relevant papers from the 1990's (Berlin et al., 1993; Mark & Prinz, 1997; Trigg et al., 1999) are case studies, analyzing the authors' experiences and observations from building and deploying repository systems. These papers focus on conventions and organizing, rather than seeking, retrieval and use. Hertzum & Pejtersen (2000) wrote about case studies of overall information seeking practices of engineers in two different organizations, but shared file repositories were mentioned only briefly. So there are no good answers to questions about information consumers' information needs and search behaviors, such as:

> Q3.2: For what purposes are shared file repositories used? What are information producers' and consumers' goals? What files in a repository are used most often, who uses them, and why?

However, there is comparable personal information management research that may shed some light on parts of those questions. Barreau & Nardi (1995) summarized findings from studies of personal electronic file organization they had each conducted separately a couple of years earlier. They described three types of information among users' personal files: ephemeral (changes often, like a to-do list), working (used frequently), and archived (used infrequently). It is conceivable that shared file repositories might at least contain working and archived information; it is less clear whether repositories would be used to store ephemeral information. Regardless, these three types are informative for thinking about what kinds of files information consumers might be searching for in a shared file repository. In a later personal information management study, Boardman & Sasse (2004) found that most files their participants mentioned retrieving fell in the "working" category. They also reported that participants sometimes mentioned accessing older items as well, "We found that although older items may be accessed erratically, they can be highly valued by people" (p. 587). From these findings it is possible to conjecture that users of shared file repositories might be likely to search for and access a mix of working and archived files, and might be more likely to use repositories that contain active, working files.

Summary

Several aspects of information consumers and repositories were mentioned in this section as potentially impacting consumers' choices about whether to seek information in a shared file repository or elsewhere. These included the consumer's awareness of the existence of the repository, his information need, and his past experiences with using the repository. The status of the files (active vs. archived) and the consumers' estimate of their potential usefulness could also play a role. Finally, the more contributions by information producers to the repository, the more benefit an information consumer is likely to gain by searching it.

PART IV: Finding

The previous section talked about why information consumers might want to use a shared file repository. This section addresses how consumers find and access information in a repository. Shared file repositories used by small groups or teams are not a known corpus, like one's own files 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 files in a shared file repository will be familiar, but most will probably be at least somewhat unfamiliar, and folders may have names that are 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 file repository. Markus (2001) wrote that information producers tend not to be very good at documenting their work. But she also argues 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 incentives 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, Crystal, Robertson, & Leadem, 2003).

The diagram below (Figure 3.3, next page), an excerpt from the concept map on page 10, shows the various influences on information consumers' search outcomes. A shared file 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 a file hierarchy explorer window. Querying requires recall, but browsing a hierarchy depends on recognition, and has been referred to as "searching without specifying" (Chang & Rice, 1993). Browsing involves visual scanning of a resource or structure, and movement through an information space, as opposed to evaluation of query results.

Just as the external representation affects an information consumer's search strategy, there is some evidence that it can also affect the structure of users' internal representations of the repository, which also influence search strategy. In collaborative situations external representations can support the negotiation of meaning and maintenance of awareness of others. External representations can even influence what information is discussed and written about in a collaborative task (Suthers & Hundhausen, 2002). These effects indicate that different forms of external representation are better at supporting certain kinds of information and activities than others. A better understanding of these effects would make it possible to design external representations so that they better support the cognition and behaviors that information consumers must employ in order to find what they need (Olson & Olson, 1999).

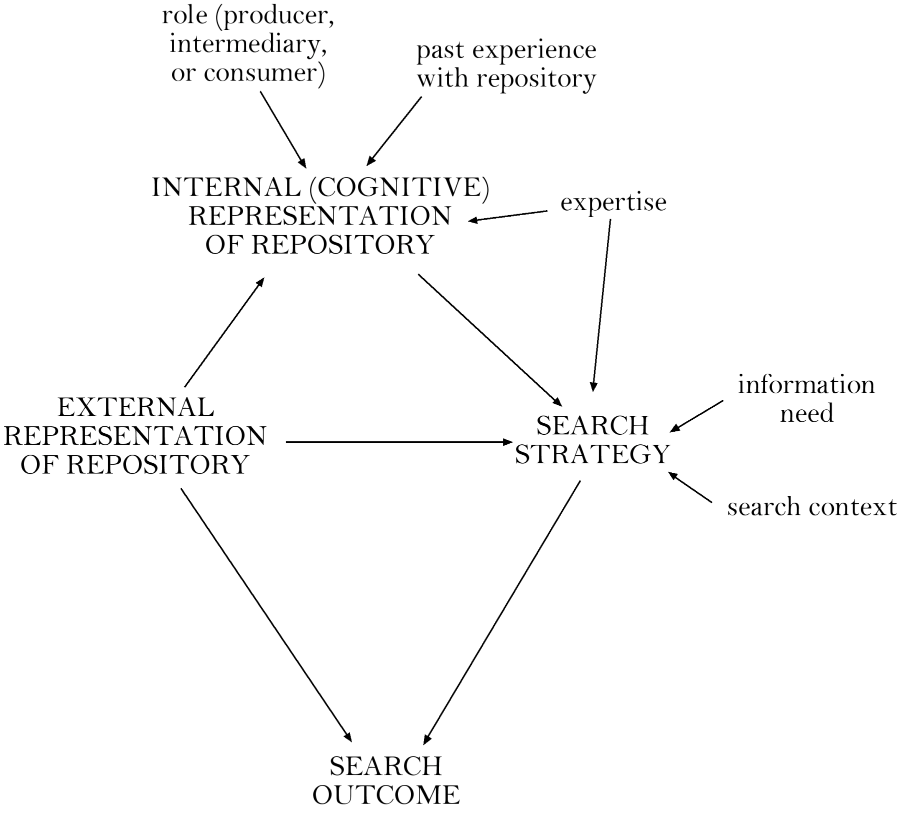
role (producer, intermediary, or consumer)
past experience with repository
INTERNAL (COGNITIVE) REPRESENTATION OF REPOSITORY
expertise
EXTERNAL REPRESENTATION OF REPOSITORY
SEARCH STRATEGY
information need
search context
SEARCH OUTCOME

Figure 3.3: Factors influencing success of search outcomes

External Representation: Information Retrieval

One form of external representation for a shared file repository is that of a traditional information retrieval system. For information retrieval, text must be represented in some way that is searchable. This can be done using several different methods: full-text, indexing, and statistical methods like vector space models, term frequency weighting, and latent semantic analysis (LSA) which statistically models the frequency of occurrence and surrounding context of words across documents (Jurafsky & Martin, 1999; Lewis & Jones, 1996). Indexing requires that documents be labeled with indexing terms either by a person, or some form of automatic document description or summarization. Dumais (2003) describes the way latent semantic analysis works: "LSA simultaneously models the relationships among documents based on their constituent words, and the relationships between words based on their usage in similar documents" (p. 493). Fully automatic statistical methods look at other words used in the document, either nearby the target word, or by counting the occurrence of words in the document, to get a better idea of what a document is really about and to distinguish documents from each other (Jurafsky & Martin, 1999). The primary advantage of LSA over other techniques is that it can determine that, for example, a document that uses the word "doctor" is similar to one that uses the word "physician", because many other words in the documents like "checkup, health, and medication" are the same. However, full-text search, as well as search of indexed documents and summarized documents is subject to word sense disambiguation problems. A search for "bass" could be for a musical instrument, or a fish.

Information retrieval systems are traditionally evaluated using two related measures: recall, and precision (Jurafsky & Martin, 1999). First, a corpus of documents must be judged by a human evaluator as to their "relevance" to a particular query. *Recall* represents the number of relevant documents returned, divided by the number of relevant documents in the corpus. It answers the question, what percentage of the relevant documents that should be returned for a particular query is the search algorithm capable of returning? *Precision* is the number of relevant documents returned, divided by the total number of documents in the corpus (both the relevant and non-relevant documents). Precision answers the question, what percentage of all documents in the corpus are related to a particular query?

A system should strive for 100% recall, meaning that the search algorithm returns all of the documents relevant to a query in the corpus. However, this could technically be obtained by returning all of the documents in the corpus regardless of whether they were related to the query. A result set with a lot of false positives – documents that are returned that are not relevant to the query – has low precision. One can imagine a user new to a particular field who is unfamiliar with the precise query terms that will get her the information she needs. Starting with general terms might return a very large, imprecise set of documents: poor precision, great recall. Precision and recall tend to have an inverse relationship such that optimizing a search algorithm for one measure usually decreases performance on the other. Precision and recall are widely used to evaluate search algorithms; however, many feel that they are impractical because in the real world (rather than a test corpus) it is not possible to obtain absolute relevance judgments for every item. Because users' information needs are dynamic, the idea that a complete set of all relevant documents for a given search can be identified a priori is simplistic (Schamber, 1994).

Bates (1998) wrote a review of research results with implications for information retrieval, that aren't usually discussed by researchers and designers in that field. She wrote that information retrieval system designers have ignored both common ground and the vocabulary problem. In her opinion:

> "Information retrieval has looked deceptively simple to generations of newcomers to the field. But IR involves language and cognitive processes based on real-world knowledge, which researchers have been trying to automate virtually since the invention of the computer." (p. 1186).

In addition, searching information stored an intranet is not the same as searching the web. Different "social forces" are behind the structure and content of an organization's intranet (Fagin et al., 2003). The Google PageRank algorithm works so well for the web because it is based on the assumption that a link to a page is like an endorsement of the content on that page. Relationships between files in a shared file repository are likely to represent different kinds of connections, such files grouped together based on an org chart, or geographic proximity which determines people who are likely to collaborate. If the links are made for these other reasons, PageRank might not produce appropriate search results. Fagin et al. (2003) analyzed the link structure of the IBM intranet, and indeed found differences from the link structures of the web. This means that ranking algorithms cannot be directly applied to an information retrieval system for a shared file repository.

Permissions are another complicated issue affecting search in a shared file repository that does not exist in other information retrieval systems. Often, information producers or intermediaries can restrict view and access rights at the level of individual files and folders. When an information consumer searches the repository, she is able to view and/or access only those files for which she has the correct permissions. This means she could have an incomplete view of the contents of the repository; the information she needs might exist in the repository and yet not be accessible. This could occur when a power structure exists within the group using the repository, and some files are restricted to only those at higher levels or with certain clearance within the organization.

External Representation: Metadata

The information retrieval methods mentioned above are limited to the text contained within the files themselves. Methods that use metadata incorporate information about how or when files were used in the past, or what is salient to the user about the files, as content tied to the files that can also be searched. Metadata is, literally, 'data about data'. It is information other than the content of a file that is associated with the file. A simple example of metadata is the information associated with a digital music file, such as artist, album, title, track number, etc.

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. Metadata can help with this, because it makes it possible to associate the details that are we already predisposed to remember with the files.

Metadata for files can be grouped into three classes:

1. One-time (automatic): information associated with the file upon its inception, like creation date and time, file type, file size, creator/owner

2. Usage history (automatic): information that can be automatically associated with a file as it is used, reflecting usage patterns, such as who has used it and date and time associated with accesses and modifications

3. User-generated (manual): tags or any other information that must be manually entered, such as the digital music metadata, above; explicit associations between items to form groups or "collections"

Automatic Metadata

Automatic metadata, including both one-time and usage history metadata, is captured without any explicit action on the part of the user. It enables the information consumer to use information other than what is contained in the text of the file when searching, without having to expend any effort to annotate or tag files. This metadata might be incorporated into a search system in a variety of ways.

Ringel, Cutrell, Dumais & Horvitz (2003) developed a prototype visualization that used metadata to help with search, capitalizing on humans' episodic memory. Memories are naturally organized into episodes with a beginning and an end in time, and relative recall of the temporal order in which things took place is fairly easy. The system they developed provided landmarks in time, in the form of event information taken from users' electronic calendars, holidays, news headlines, and digital photos, and used date and time metadata to display files and emails next to a timeline containing these landmarks. Results of a user study they conducted indicated that searches were completed faster with the landmark interface than with an interface that showed only dates. However, relative timing might not be all that helpful in situations like shared file repositories, where an information consumer has little direct personal experience with the content they search for. It is not reasonable to expect a newcomer to the group, for example, to remember the relative timing among events and files.

Metadata associated with files can also be used to detect behavior patterns and infer preferences, in addition to assisting with search and retrieval. For example, recently used files are more likely to be used again; a record of who has been using the files might help someone else locate what they are looking for. This information can be combined to infer usage patterns for the repository, and build profiles of individual users' interactions with the repository. Teevan, Dumais, & Horvitz (2005) proposed incorporating usage history metadata, such as web pages viewed, documents opened, and emails sent and received into models of users' short term and long term interests. They suggested using these models to create a personalized ranking of search results for a given user, essentially inferring what he might be interested in based on what he has viewed in the past. While this is an interesting idea, in tests using a scoring method developed by the researchers, the personalized ranking algorithm performed significantly worse than Google's ranking of the same search results, indicating that careful choices need to be made about what should be incorporated into the model and how various aspects should be weighted. There is reason to believe this technique might work better for shared file repositories than for the web; usage history can be collected for the entire corpus of files being searched. This additional information about file usage as well as user behavior may enable the system to make more accurate predictions.

User-generated Metadata

A system that incorporates user-generated metadata provides the information producer, intermediary or consumer with the ability to annotate files with additional labels or keywords they generate themselves, also called tags. User-generated metadata can include information that it is not possible to capture using automatic methods, such as perceived relationships among files, or what the file is "about" in a particular user's mind. These tags are intended to make the file easier to retrieve later. The idea of assigning user-generated metadata to files sounds very similar to the formal categorization and indexing process undertaken by library professionals. In fact, a standard has been created called the "Dublin Core metadata schema" (`http://www.dublincore.org/documents/dcmi-terms/`) for users to follow when annotating content with metadata. The Dublin Core defines 15 basic "elements" to be completed about every file, eliminating the need to employ professionals to do this task. However, the schema can be difficult to use for authors who have not been trained as catalogers (Greenberg et al., 2003). For example, the elements "description" and "subject" might seem very similar to someone with little cataloging experience. The "description" element is defined as "an account of the content of the resource," whereas the "subject" element is "the topic of the content".

Often, file authors see no point in providing metadata at all. In a study of a tool they had developed to support the task of metadata generation by document authors, Greenberg et al. (2003) wrote:

> "One [author] noted skeptically that metadata creation 'seems like just one more thing added to all the other 'one more things' that eventually take up all your time.' Another put it bluntly: 'not my job or responsibility or interest,'" (p. 6).

These comments reflect a mismatch of incentives for metadata creation, similar to that discussed in Part I (pp. 13-14). The users who spend time and effort contributing metadata are often not the same as those who would benefit most from an annotated corpus. Only those few information producers who might be intrinsically motivated to spend time and effort contributing metadata would do so. Therefore, the metadata for a corpus would be incomplete, and any searches using metadata would return only partial results.

In the context of shared file repositories and so-called "social bookmarking" systems like `flicker.com` and `del.icio.us`, proponents of user-generated metadata (hereafter called "tags") argue that in contrast to the structured Dublin Core schema, the biggest advantage of tagging is that it is unstructured. There are no rules to follow, meaning little effort or overhead is required to adhere to a standard when creating tags (Marlow, Naaman, boyd, & Davis, 2006). But, this might mean that the effort is offloaded to the time of search and retrieval of the items that have been tagged. If I expend very little effort now to choose tags that might be meaningful to me in the future, I might choose poorly or inconsistently, or be unable to imagine the information that will be salient to me (or others) at retrieval time. This could result in difficulty finding the item again later. Hastily-selected tags might be even more of a problem in situations where information producers and consumers are not the same person. Also, it seems that the vocabulary problem would apply equally to filenames and tags, and that the degree of common ground would again be an important factor in how useful the tags are.

> Q4.1: Is tagging more like labeling or assigning something to a category? What are the implications of this for shared file repositories?
>
> Q4.2: How does the amount of effort expended to select "good" tags affect search and retrieval of tagged items?

Some researchers believe user-generated metadata in the form of tags is at its most useful when both files and tags are publicly available to all users of a collaborative system. This has been called "social tagging", and a list of social tagging systems is presented in Marlow et al. (2006), including `del.icio.us`, `flicker.com`, `CiteULike.org`, and `YouTube.com`. Marlow et al. suggested that social tagging is an implementation of the "strategy of unlimited aliasing", which Furnas et al. (1983) concluded through statistical analysis might be a solution to the vocabulary problem (described on page 17 of this paper). Furnas et al. wrote that a system capable of remembering every word anyone has ever used to refer to an object (file) could achieve retrieval success rates approaching 100%, if 15 or more words are stored for each object. The more unique words stored for each object, the more likely a user would be to select a word already associated with the object, even if the word had only been used by one other person. In a test of this hypothesis, Gomez, Lochbaum, & Landauer (1990) asked 60 women with little computer experience to create labels for recipes, instructing them to choose words that would be helpful to other people trying to locate the recipe. They found that when the system's index included all of the labels generated by the women, people were able to find 76% of the recipes they were tasked with finding, and only 25% when the index included only a few of the more common labels.

The "strategy of unlimited aliasing" is exactly what a social tagging system like `del.icio.us` does. Users of `del.icio.us` are able to store bookmarks online, and select up to 10 different tags for each web page they bookmark. One might expect that web pages bookmarked and tagged in `del.icio.us` with a large, diverse set of words would be easier for a random user to find using tag queries, than web pages tagged with only a few words. Social tagging would likely be less beneficial for a shared file repository serving a small group of users, than a large group. Small groups are probably not of sufficient size to produce enough tag diversity to eliminate the vocabulary problem. However, one exception to this might be situations where groups share a high degree of common ground. The vocabulary problem states that two random people choose the same word for the same object less than 20% of the time. Common ground might increase this percentage significantly, meaning that fewer unique tags would be required to achieve high search success rates.

> Q4.3: Are social tagging systems a solution to the vocabulary problem for search? How many users are necessary for this solution to be viable? What are the implications of common ground for social tagging systems?

External Representation: Hierarchy vs. Search

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. There are other places besides the architecture of a building where information can be encoded, such as processes and procedures, technologies and equipment, even the internal layout of an office or workplace (Argote, 1999). Knowledge that has been "stored" this way is unlikely to disappear from an organization when any particular individual leaves.

Hierarchies may convey information about the structure and content of a shared file repository that information consumers would be unable to access if they were to interact with the repository using a query 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, Phuwanartnurak, Gill, & Bruce (2005) found that folder hierarchies and filenames provide meaningful information that helps people summarize content as well as organize it. Grouping things manually allows for the formation of visible relationships between files. 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, Robbins, Dumais, & Sarin, 2006).

There have been few empirical evaluations comparing different external representations for search. Quan, Bakshi, Huynh, & Karger (2003) conducted a user study with 21 MIT CS grad students who categorized and then tagged a corpus of 60 `ZDNet.com` news stories. After a week's delay the participants returned to the lab and completed search tasks using both hierarchy and tags. In one interface, participants directly navigated the hierarchy. In the other, they used tags as filters, rather than search terms; i.e., all tags were visible on the screen, and clicking the checkbox next to a tag filtered out news stories that had not been tagged with that word. Participants were 17% faster at completing the search task when they used tags vs. categories; however, in their paper Quan et al. excluded searches where participants "gave up" because they couldn't find the item, and did not report how the analyses looked with and without these data. There was no significant difference between the two conditions on how many users "gave up".

> Q4.4: How do different external representations (full-text search, metadata+full-text, or navigating/orienteering a hierarchy) affect the structure of users' internal representations for a shared file repository?

When navigating 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 shared file repository hierarchy is similar in some ways to a website with a link structure: folder names are like link text. An information consumer is able to browse until she recognizes something related to what she is looking for (Bruce, Jones, & Dumais, 2004; Trigg et al., 1999). In a study conducted by Boardman & Sasse (2004) users searching their personal repositories used a combination of browsing and sorting of folders. Because they were searching their own files, they exhibited a tendency to know approximately where in the hierarchy to start looking. From there they used recognition memory navigate to the particular file they wanted. Teevan, Alvarado, Ackerman, & Karger (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 files, but could not explicitly recall the path or any of the folder names for where it was stored, making it very difficult for her to search for the item using a query interface. Orienteering allowed the participant to find the file, because the information she needed at each step to prompt her next step via recognition was built into the hierarchy structure. All she had to do was be able to recognize the next step, not recall it.

Internal Representation

The previous section described, compared, and discussed implications of different kinds of external representations. Below I will focus on shared file repository users' internal representations, referred to above as mentions of 'memory' or 'recognition'. The success of an information consumer's search depends upon an interaction between his internal representation of the information contained within the repository, and his interpretation of the external representation with which he interacts.

Categorization is a cognitive shortcut that allows us to predict, infer, and assume facts and relationships among things we encounter in our daily lives. Psychologists have been able to uncover many aspects of mental categories over the years, but there is still no consensus about the underlying cognitive mechanisms that allow us to form categories (Murphy & Lassaline, 1997). Categories to some extent reflect regularities or structure in the nature of what's being categorized; they are not arbitrary (Mervis & Rosch, 1981). There is also a "basic level" of categorization that people seem to prefer. The basic level is specific enough to distinguish members of the category from other basic level categories (i.e., cat vs. dog), but not so specific that one must bring to mind specifics that are unnecessary for telling the categories apart (i.e., labradoodle vs. siamese). The basic level of categorization is cognitively efficient because it maximizes within-category similarity relative to between-category similarity. According to Bates (1998), there has not been an investigation into how well formal classification systems correspond with basic level categories. This comparison has not yet been undertaken for folders in personal or shared file repositories, either.

> Q4.5: Do basic-level categories appear in personal or shared file repositories? What is their effect on how the repositories are used?

Not all members of a category are equally representative of a category; "gradients of representativeness" exist, and people are generally able to judge whether a particular exemplar is typical or atypical of a category (Mervis & Rosch, 1981). The catch is that these judgments tend not to be consistent for most categories, both within and between subjects. In a shared file repository, if a particular file is not a very good match with any of the existing folders (categories), one might expect that it could end up in many different folders, and there would be a lot of variability in the locations information producers would choose. This inconsistency in category assignment would make it unlikely that an information consumer would look for the file in the correct place.

It is possible to discover the structure of users' internal representations of the content in a shared file repository, through knowledge elicitation methods. The organization of hierarchical mental categories can be elicited via a free recall task (Reitman & Rueter, 1980). Pauses occur during the free recall that correspond with separations among "chunks" of information in memory, and by asking subjects to perform repeated free recall trials starting from different points, a hierarchical structure can be deduced. Other techniques are useful as well, including "conceptual" techniques like triad tests, and "process tracing" techniques such as think aloud protocols and other verbal reports. Cooke (1994) recommends using a conceptual knowledge elicitation technique in conjunction with a process tracing technique to triangulate as closely as possible the true internal representation.

Users' interactions with shared file repositories can affect their internal representations for the structure of the repository (McCreadie & Rice, 1999). Searching or browsing (orienteering) a shared file repository is a distributed cognitive task (Zhang & Norman, 1994), where part of the information that is required for task completion is present in the head of the user, and part is in the repository. According to Zhang & Norman, external representations not only serve as memory aids; they can structure or constrain behavior and cognitive processes. It might be that searching via a query interface relies more on an information consumer's internal representation, than does orienteering via a hierarchy structure that presents the relationships among the concepts clearly. Also, interacting with a structured hierarchy might have different implications for the structure of the internal representation than orienteering via tags, or searching using a query interface.

> Q4.6: How do information consumers' internal representations for a shared file repository correspond with the external representation?

Search Strategy

When an information consumer has decided to search a shared file repository, she is then faced with the mass of information that has been organized according to the idiosyncratic strategies of other members of her workgroup. She must somehow be able to decide what is relevant to her information need, and what is not. Relevance is a relationship between the user's interpretation of the information in front of her and her search context, which includes her goals, needs and assumptions at that moment (Harter, 1992). This means relevance is subjective, and situational. Relevance has historically been used as a measure of an information retrieval system's success, as in "does the system return enough relevant documents." But many information behavior researchers now agree that it is very difficult, if not impossible, to measure absolute relevance for all people, all the time (Schamber, 1994). Users are able to make personal judgments of relevance throughout their information seeking activities; this is what guides them through the shared file repository or whatever the corpus is at hand.

Once an information consumer has decided that useful information is likely to be present in the repository, he must wade through the contents or search results and make judgments about what is relevant and what is not. These judgments might be made more difficult in shared file 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 & Pejtersen, 2000). Having access to the files does not mean access to the meaning and implications behind those files, 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.

> Q4.7: How do aspects of a file's history or context interact with the external representation format of the repository to affect an information consumer's relevance judgments and search process?

The difficulty of this task is compounded by the fact that much of an organization's knowledge is stored in the workers themselves, or externalized in places like processes or technologies rather than packaged for storage and re-use in shared file repositories (Argote, 1999). People with adequate background and experience are necessary to translate between all of the different places where this information is stored, and use it appropriately (Ackerman & Halverson, 2004). Grudin (2006) made a prediction that increases in computer processing power and storage capacity will enable information that has in the past been captured in other places to be stored and accessed via computer technology. However, search and organization technologies will have to improve a great deal in order for any of that information to be accessible or understandable by someone without all of the background knowledge and context.

Prototypes

Several research prototypes have been developed by various groups in the past 10 years or so, using many of the ideas discussed in this section. These prototypes are hypotheses of a sort, guesses made by researchers and designers about the relationship artifacts, cognition, and behavior (Woods, 1998). The research prototypes discussed in this section all use metadata in one way or another, and in most cases combine it with text processing and analysis techniques to improve an individual's retrieval of files. (A notable exception is the Prinz & Zaman (2005) Semantic Workspace Organizer, which makes suggestions for where a file should be stored in a shared file repository based on text analysis of the document to be filed, the documents in the repository, and a profile of the user.) The design goals of these systems seem to fall into two camps: those trying to eliminate files-and-folders altogether and come up with a new metaphor for interacting with files, like Lifestreams and Haystack; and those trying to improve search for personal information management like Stuff I've Seen and Phlat. Details about each of the systems discussed in this section can be found in Table 2 starting on page 40. These systems are discussed here because they incorporate various concepts and ideas that have appeared so far in this section.

In these systems, most files have metadata associated with them, sometimes called "attributes" or "properties". Systems differ in whether they allow the user to manually make changes to automatically generated metadata, or create groups or "collections" by adding tags (essentially tagging) rather than just by querying on some specific attribute-value pair and retrieving a group of search results. Leaving rigid hierarchies behind means that access to items is no longer tied to "where" they are stored, therefore it requires only a few steps to reorganize a collection of documents depending on the task or purpose at hand (Dourish, Edwards, LaMarca, & Salisbury, 1999).

However, the architectures of the systems differ, and so do the user interfaces. In order for metadata to be searchable as well as full text, a database containing all the content on a user's computer including the metadata must be created, as with the Haystack system developed as part of project Oxygen at MIT (Karger, Bakshi, Huynh, Quan, & Sinha, 2005). Or, the metadata has to be built into the files themselves and be accessible to the operating system, via an index that must be kept updated, as with "Stuff I've Seen" (Dumais et al., 2003) and Phlat (Cutrell et al., 2006), both developed by Microsoft Research. The user interfaces appear to be in different stages of development. Most incorporate some kind of querying and perusal of results. This takes different forms and has varying levels of feedback and interactivity for refining the search. For example, an interface was developed for Phlat (Cutrell et al., 2006) that "merge[s] search and browsing"; all queries seem to the user to be essentially filters. The difference between a query and a filter in Phlat is that the user chooses a term and types it in as a query; possible things to filter on are subsequently provided as a list of checkboxes in the interface based on the metadata properties of the items returned. Cutrell et al. present log data describing how people at Microsoft used the system. Users entered queries that averaged 1.60 words long. Forty-seven percent of all queries involved a filter, and 1/3 of those queries involved multiple filters. Seventeen percent of all queries used filters only, no typed-in text.

(Chalmers, 2002) describes "Recer", a prototype recommender system that suggests URL's and filenames that seem related to a user's current activities. This system is relevant for a discussion of shared file repositories, because the recommendations are based on the metadata of a group of people, not an individual. The prototype logs everyone's activity, including what files are opened and URL's accessed, and when, and some of the text of these items. Then, if the system detects a relationship between what one user is currently looking at and what others have done in the past, it recommends that the user take a look at those same things.

These metadata-based systems seem to solve many problems mentioned in this paper regarding decisions about where to put things and having to remember rules for folders. They offer the possibility of being able to search based on whatever a user happens to remember about what they are looking for. However, this positive is also a negative. The user *must* remember or recognize something about the file she is trying to find! The content in these systems can be accessed via a number of different "views" that can only be created if the user has some filter to apply; it is not possible to navigate or orienteer towards something without metadata. For an individual user and their own content, this is inconsequential, but when information producers and consumers are not the same person it can be a real problem. How is one to remember or recognize something about a file one has never seen before?

One thing these systems have in common is that evaluations with users are rarely reported. None of the papers cited here have put their prototypes to a fair, head-to-head test against each other, or the standard desktop metaphor for personal information management, to see whether or not the improvements they expect really do materialize. To be sure, this is a hard thing to do. A test like this requires a robust, nearly production-ready prototype, with a user's own files incorporated and indexed, and use of the system over time so that true behavior patterns can develop.

> Q4.8: What factors influence the performance of different external representations (i.e. full-text search, metadata plus full-text, or navigating/orienteering a hierarchy) for finding information in a shared file repository? How does this depend upon the purpose and context within which the repository is used?

Summary

This section has discussed the interactions among internal and external representations of a shared file repository, factors that affect information consumers' search strategies, and how these three elements affect search outcomes. Different forms of external representations were introduced, including information retrieval via querying full text, text that has been summarized and indexed, and metadata. Automatic and user-generated metadata are two kinds of contextual information that can be added to files. The more context an information consumer knows or can access for the files in the shared file repository, the better able he is to assess the relevance of the files for his information need.

SYSTEM	OVERVIEW	SEARCHING	UI
Lifestreams (Freeman & Gelernter, 1996)	Alternative to file-and-folder hierarchy "desktop metaphor". Eliminates need to name files or use directories. All documents are ordered chronologically.	Streams can be searched by specifying a timeframe and browsing, or by full-text query. Query returns a temporary result set called a "view", which is a subset ordered by time. No mention of ability to edit metadata.	Interface displays a chronologically ordered "stream" of all documents. Thumbnails are displayed on screen to summarize a "stream" visually.
Presto (Dourish et al., 1999)	Improve upon rigid file-and-folder hierarchy as a way to organize documents. Documents have metadata called "attributes" that are automatically captured; users can also create their own attributes.	Documents can be grouped by any combination of attributes into "collections", via queries. Documents can be added to collections manually, updating the "inclusion list" for the collection, not the document's metadata. Documents can exist in multiple collections.	Collections are displayed graphically on the desktop, and changes to the query result in visual feedback as the collection updates.
Recer (Chalmers, 2002)	Recommends URLs and files that the system determines are relevant to current user activity. Activity of a group of users is tracked, and a time-ordered history for each user is maintained.	Queries happen implicitly when a web page is loaded or a file is accessed. Recommendations are based on frequency of use and co-occurrence with other items. Users do not have direct access to manipulate metadata.	Not described.
Stuff I've Seen (Dumais et al., 2003)	Goal is to help people find information on their personal computer. Combines documents, email, and web pages into one index.	Queries operate on an index of full text and metadata. Metadata is all automatic, no user-specified attributes or values.	No explicit "submit" button for queries. Checkboxes act as filters for the different values for the metadata types (i.e. type of item: outlook, file, web page). Query box, filters, results displayed on one screen.

Table 2: A comparison of personal information management research prototype systems (part 1)

SYSTEM	OVERVIEW	SEARCHING	UI
Memory Landmarks (Ringel et al., 2003)	Focuses on time metadata in order to take advantage of the chronological and episodic nature of human memory.	Search by entering text queries, which returned a result set displayed in chronological order with the addition of "landmarks" (see UI description, right)	Interactive visualization combining time metadata with "public landmarks" – holidays, news headlines, calendar events, digital photos – "anchors" to help with recall of things that happened at the same time. UI built to work with SIS (above).
Semantic Workspace Organizer (SWO) (Prinz & Zaman, 2005)	Users have trouble finding items in shared workspace systems; the SWO makes recommendations for where things should be stored based on activity history metadata and text analysis profile of items in each folder.	Searching not covered in the paper. Activity history metadata are stored for each document, and profiles of folders are created via text analysis of items in folders. Intended for groups, not individual users.	UI displays the file-and-folder hierarchy of the system in the left pane, and suggested locations for uploaded file in the right pane. The user can select any location as the destination folder.
Haystack (Karger et al., 2005)	All users have different information needs and preferences; systems should be flexible rather than hard-coded and rigid. Stores objects (documents, email, web pages, etc. – anything) and their metadata (called "properties") to provide flexibility in as many ways as possible.	Queries of full text and metadata, or "properties", that can describe documents, or relationships between documents. Objects can be part of multiple "collections". Manual updates to metadata are possible.	Interaction is via metadata-based "faceted" browsing, or "orienteering" from a starting object. Browsing advisor suggests "similar" items based on properties they have in common.
Phlat (Cutrell et al., 2006)	Goal is to take advantage of contextual memory surrounding objects one might search for, to help people find information on their personal computer. Provides as much immediate feedback as possible in the UI.	Merges search and browsing, and queries indexed full-text and property-value (metadata) pairs so that users can search on whatever they happen to remember. Also supports addition of user-generated tags.	Distinction between queries typed into a text box, and property value filters applied via check boxes. Active queries and filters are visible at all times, along with the number of items of each property value that were returned by the active query/filter combination.

Table 2: A comparison of personal information management research prototype systems (part 2)

Searching a shared file repository is a distributed cognitive task, and the external representation affects both the internal representation and search behavior of an information consumer. Hierarchies make visible information about the structure and content of a shared file repository that information consumers would be unable to access if they were to interact with the repository using only a query interface. Searching via querying requires different strategies and cognitive processes (recall) than searching via orienteering or browsing a static hierarchy (recognition). It is not clear how the external representation affects the internal one, or vice versa, how either affects the search strategy; however, it is possible to elicit the internal representation and compare it to the external representation.

Finally, several research prototypes were described and compared. Most have been developed for personal information management and are based on assumptions that might not hold for shared file repositories, such as the assumption that the repository user has seen the files it contains before. Also, the prototypes have not undergone rigorous testing to find out whether they are actually an improvement over the standard desktop metaphor for organizing and managing information.

RESEARCH QUESTIONS AND DESIGNS

The practice of sharing files via an online repository presents an opportunity to conduct research about a problem that affects many people, in an area where there are more open questions than published answers. In this paper I have described factors that affect interactions with shared file repositories and suggested many research questions (a complete list of questions can be found in Appendix A, starting on page 53). The questions in Part I: Storing have to do with factors that affect information producers' decisions to contribute to shared file repositories. Part II: Organizing focuses on how the behavior and cognition of information producers is cumulatively externalized into the structure, or external representation, of the repository. In Part III: Seeking, the choices of information consumers regarding whether to look for the information they need in the repository are considered, as well as what they might be looking for. Finally, in Part IV: Finding, factors affecting the search outcomes of information consumers are considered, including the two-way interaction between the external and internal cognitive representations of the repository. With so much research to be done, where does one start, and how does one prioritize?

The difficulty individual users have with organizing and finding their own files is the subject of personal information management (PIM) research. These same problems are also present in shared file repositories; however, PIM research results and designs are based on assumptions that do not apply in collaborative settings, and therefore cannot be applied directly to shared file repositories. Shared file repository tools today are immature, essentially just personal information management tools, with support for permissions and in some cases version control and allowing multiple people to use them simultaneously. The problems with these tools are severe enough that in some instances users circumvent them altogether – the title of this paper is reflective of this. The tools do not have the right functionality to support user needs and goals. Until problems of this nature are solved, it is very difficult to answer questions regarding motivation, incentives for contribution and use, and adoption that require studying viable repositories with the right functionality, good usability, and a solid user base. The questions in Part II: Organizing and Part IV: Finding ask how theory from other fields can be applied to the design of shared file repository tools having the right functionality. A design is a prediction about the characteristics of a useful and usable artifact for a given set of circumstances. It encodes the designer's understanding of and assumptions about user goals and cognition and usage context into an external form with which users can interact. Answers to questions about language use and common ground, memory and representation, and social context will provide a knowledge base from which the next generation of shared file repository tools can be designed.

To this end, I selected the following research questions for further discussion here:

> Q2.3: How does common ground affect information producers' choices of labels for files?
>
> Q2.4: To what extent do information producers "package" their contributions to a shared file repository? How does this affect repository use by information consumers?

Problems with finding files in a shared file repository start with the creation and addition of files to the repository, and Question 2.3 asks whether common ground can help to explain variations in the structure of shared file repositories. Question 2.4 takes this a step further to ask whether information producers in the real world take common ground into account when contributing to a repository, and whether this makes any difference for information consumers who search the repository.

Research Design One: Experiment

Q2.3: How does common ground affect information producers' choices of labels for files?

This research design is intended to examine one factor that might affect the external representation of a shared file repository, and establish a causal relationship between that factor and characteristics of the repository. It has been shown that people tailor the language they use in conversation for their conversation partner or presumed audience (Krauss & Fussell, 1991; Russell & Schober, 1999; Schober & Clark, 1989). This experiment will investigate whether this is true in a situation involving language use that is not a conversation – that of selecting labels and tags for files that might be stored in a shared file repository.

Fussell & Krauss (1989) presented line drawings to participants, and asked them to create short descriptions of the drawings either for themselves to use to identify the drawings later, or for someone else to use in the same task. They found differences in length of description depending on whether it was written for self or others. Then, some time later participants returned to the lab and were assigned to one of three conditions: people who used their own descriptions to identify the drawings, people who used another participant's descriptions written for others, and finally people who used another participant's descriptions that were written for self. Participants using their own descriptions performed the best, and people using others' descriptions that were written for self performed the worst.

Using their study as a model, this experiment will be a 3x2 mixed design. Perceived audience will be between subjects, and have three levels: self, familiar other, unfamiliar other. A familiar other might be someone in the same lab group or project team, while an unfamiliar other might be a potential new collaborator from an overseas office. Annotation format will be within subject and have two levels: label or tags. The within subjects condition will be counterbalanced such that some participants generate labels first, and some generate tags. A label is equivalent to a filename, a unique identifier for a file. A tag is like a keyword, and multiple tags can be assigned. The null hypothesis is that intended audience has no impact on the external representation of a shared file repository or search task performance, and that annotation types perform equally in search tasks. The experimental hypothesis is that annotations created for familiar others outperforms the other conditions, and that tags outperform labels.

The procedure of the experiment will be as follows: in part 1, participants in each of the three "audience" conditions will come to the lab and create both labels and tags for a set of files. Then several weeks later, they will return for part 2 of the experiment, and match the labels and tags that they are given to the appropriate files. Each participant will be asked to match only a subset of the files, and they will not be given labels and tags for the same files. The labels and tags they receive will be one of the following:

- their own annotations written for self
- their own annotations written for familiar others
- their own annotations written for unfamiliar others
- others' annotations written for self
- others' annotations written for familiar others
- others' annotations written for unfamiliar others

Two analyses will be conducted: the tags and labels generated in part 1 will be examined for differences between the three "audience type" conditions, and performance in part 2 will be assessed.

Research Design Two: Verbal Protocols and a Survey

Q2.4: To what extent do information producers "package" their contributions to a shared file repository? How does this affect repository use by information consumers?

The experiment described above will provide results indicating whether or not intended audience can have an effect in the lab on the filenames and user-generated metadata such as tags. However, the experiment tells us nothing about whether or not users in the real world ever consider others when generating filenames and tags (i.e., "packaging" them).

Questions about how people make decisions about filenames and locations have been investigated for personal information management (for example: Barreau, 1995). Most of these have been case studies involving small sample sizes, and descriptions of behavior after-the-fact. A common technique used in these studies is to ask users to give the researcher a "tour" of their personal repository or their physical office, talking about the various places, systems, and strategies they use for organizing and finding information. I do not wish to replicate these studies in yet another setting.

Instead, collecting verbal protocols from shared file repository information producers as they make choices about where to store files, what to call them, and perhaps how to tag them, is a way to capture data about factors that affect these choices as people are making them, rather than data about what they remember thinking about or being important at that time. People are notoriously bad at reconstructing past events accurately, and instead produce an idealized version of events that represents what they believe they do, rather than what they actually do. For this reason, interviewing them about their behavior is less ideal than capturing actual behavior. Participants for this experiment should therefore be shared file repository users; this experiment will require identification of several suitable candidate repositories from which to recruit. It will also be necessary to classify users into 'producer' vs. 'consumer' categories – this would likely be done via a questionnaire early in the study.

The reason for collecting verbal protocols rather than simply observing, is to attempt to capture aspects of the processes employed by information producers when adding files to a repository. It is impossible to infer this kind of information by simply watching users file things. I want to know things like, what information about group members, both information producers and consumers, is salient? What other things do users consider? What tradeoffs must be made? Does common ground play a role in filenames, labels, and locations? Participants will be asked to think aloud while adding files to the repository. Their responses, the verbal protocols, will be transcribed and analyzed.

Verbal protocol analysis consists of developing a coding scheme and applying it to cleaned and segmented protocols (Chi, 1997). From there a code-and-count procedure might be followed, but in this case the goal is to describe user behavior and correlate it with the structure of external representations of shared file repositories, rather than compare two different groups that vary only according to experimentally controlled variables. So instead, coded verbal protocols will be used to create a process model for how information producers make these decisions. One drawback of using verbal protocol analysis is that not all researchers agree that the information in the protocols is an accurate representation of what is going on inside people's heads (Ericsson & Simon, 1980). However, for this purpose, these data cannot be collected any other way.

The research question at the beginning of this section also asks about possible effects "packaging" might have for information consumers who use the shared file repository. It might be best to try to obtain answers to this question via a large-scale survey of shared file repository users. In this way, repositories in which "packaging" generally takes place might be identified (question design can be informed by the results from the verbal protocol analysis), and respondents can also be asked about outcomes. The data can then be analyzed using multivariate techniques to see whether any relationships between the two are present.

SUMMARY AND CONCLUSION

This paper has described factors affecting how shared file repository users behave with respect to the repository. Files in such online storage spaces are shared among members of workgroups, where “shared” means “have in common” rather than “send and receive”. Shared file repositories can contain important, mission-critical information, and yet accumulate content over time and become poorly organized such that users have difficulty finding the files they need. It is reasonable to expect a user of a personal repository to have at some time or another seen or used the files within it; many systems for personal information management are designed based on this assumption. However, the assumption does not hold for shared file repositories, which are collaborative systems. A given user of a shared file repository can expect to be familiar with only some of the contents of the repository, some of the time, because other group members also have access and can create, rename, move or delete files as necessary.

Workgroup members each choose whether or not they will use their group’s repository, or share files via some other means. In the case of information producers, their choices about whether or not to contribute files to the repository may depend upon group- and organization-level factors such as characteristics of their workgroup, organizational culture, and incentives. Once they have chosen to contribute, they must label the files they have elected to add to the repository, and choose a location or tags by which the file will be accessible to others. Naming conventions are very hard to keep consistent, especially among a group of individuals who each have their own unique way of doing things. Locations, too, are affected by idiosyncratic strategies, expertise of the information producer, and salient aspects of a file and its context at the time it is being stored. Inconsistency and lack of agreement on how files should be named and where they should be stored makes it very difficult for an information consumer to locate a file contributed to the repository by someone else. However, common ground shared by group members may mitigate this problem, such that those information producers who take their group members into consideration may create repositories that are easier for others to search.

Information consumers also make choices, about whether to seek information in a shared file repository or elsewhere such as via direct email from a colleague. Their choices are influenced by the their information needs, how much effort they perceive will be required to search the repository, and how useful or valuable they consider the files in the repository to be. When they have chosen to search the repository, many factors interact to contribute to the search outcomes. Searching a shared file repository is a distributed cognitive task, meaning that some of the information needed to complete the task exists in the external representation (the repository), and some of it resides in the internal representation (the user’s knowledge). The contents of a shared file repository can be accessed and interacted with in several different ways (i.e., hierarchy/browse vs. query/search interface); these different external representations affect information consumers’ search strategies and internal representations for the contents and structure of the repository. Properties of certain forms of external representation may be more appropriate for shared file repositories than others. For instance, query interfaces rely on users’ recall of the text they search for, whereas browsing or “orienteering” a hierarchy or metadata-based groupings of files requires only recognition or a hint of the “scent” of the correct path. Because not all information consumers can be expected to have previous experience with the information they search for in a shared file repository, external representations geared toward orienteering may perform best.

Throughout this paper, many open research questions were highlighted. Several research designs were outlined for some of these questions, having to do with common ground and representation. Future answers to these questions will be a first step towards designing the next generation of shared file repository tools, so users will no longer say "Just email it to me!"

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.

Ackerman, M. S., & Halverson, C. (2004). Organizational Memory as Objects, Processes, and Trajectories: An Examination of Organizational Memory in Use. *Computer Supported Cooperative Work (CSCW), 13*(2), 155 - 189.

Argote, L. (1999). Organizational Memory. In *Organizational Learning: Creating, Retaining, and Transferring Knowledge* (pp. 67-97): Kluwer Academic Publishers.

Barreau, D. (1995). Context as a factor in personal information management systems. *Journal of the American Society for Information Science, 46*(5), 327-339.

Barreau, D., & Nardi, B. A. (1995). Finding and reminding: File organization from the desktop. *SIGCHI Bulletin, 27*(3), 39-45.

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., & 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., & Wharton, C. (1993). *Where did you put it? Issues in the design and use of a group memory.* In the Proceedings of SIGCHI conference on Human factors in computing systems, Amsterdam, The Netherlands.

Blair, D. C. (2002). Information Retrieval and the Philosophy of Language. In B. Cronin (Ed.), *Annual Review of Information Science and Technology* (Vol. 37, pp. 3-50). Medford, NJ: The American Society for Information Science and Technology.

Blair, D. C., & Kimbrough, S. O. (2002). Exemplary documents: a foundation for information retrieval design. *Information Processing and Management, 38*, 363-379.

Boardman, R., & Sasse, M. A. (2004). *"Stuff goes into the computer and doesn't come out": A cross-tool study of personal information management.* In the Proceedings of SIGCHI Conference on Human Factors in Computing Systems, Vienna, Austria.

Bruce, H., Jones, W., & Dumais, S. (2004). *Keeping and Re-Finding Information on the Web: What Do People Do And What Do They Need?* In the Proceedings of ASIST 2004: Proceedings of the 67th ASIST annual meeting, Chicago, IL.

Chalmers, M. (2002). Awareness, Representation and Interpretation. *Computer Supported Cooperative Work (CSCW), 11*(3-4), 389 - 409.

Chalmers, M. (2003). Informatics, Architecture and Language. In K. Hook, D. Benyon & A. J. Munro (Eds.), *Designing Information Spaces: The Social Navigation Approach* (pp. 315-342). London: Springer.

Chang, S.-J., & Rice, R. E. (1993). Browsing: A multidimensional framework. *Annual Review of Information Science and Technology, 28*, 231-271.

Chi, M. T. H. (1997). Quantifying Qualitative Analyses of Verbal Data: A Practical Guide. *The Journal of the Learning Sciences, 6*(3), 271-315.

Clark, H. H. (1996). Common Ground. In *Using Language.* Cambridge: Cambridge University Press.

Constant, D., Kiesler, S., & Sproull, L. (1994). What's mine is ours, or is it? A study of attitudes about information sharing. *Information Systems Research, 5*, 400-421.

Cooke, N. J. (1994). Varieties of knowledge elicitation techniques. *International journal of Human-Computer Studies, 41*(6), 801-849.

Cutrell, E., Robbins, D., Dumais, S., & Sarin, R. (2006). *Fast, flexible filtering with phlat.* In the Proceedings of Proceedings of the SIGCHI conference on Human Factors in computing systems, Montreal, Quebec, Canada.

Dourish, P. (2004). What we talk about when we talk about context. *Personal and Ubiquitous Computing, 8*(1), 19-30.

Dourish, P., Edwards, W. K., LaMarca, A., & Salisbury, M. (1999). Presto: an experimental architecture for fluid interactive document spaces. *ACM Transactions on Computer Human Interaction, 6*(2), 133-161.

Dumais, S. (2003). Data-driven approaches to information access. *Cognitive Science, 27*(3), 491-524.

Dumais, S., Cutrell, E., Cadiz, J., Jancke, G., Sarin, R., & Robbins, D. C. (2003, July 28 – August 1, 2003). *Stuff I've Seen: A System for Personal Information Retrieval and Re-Use.* In the Proceedings of SIGIR'03, Toronto, Canada.

Ericsson, K. A., & Simon, H. A. (1980). Verbal Reports as Data. *Psychological Review, 87*(3), 215-251.

Fagin, R., Kumar, R., McCurley, K., Novak, J., Sivakumar, D., Tomlin, J. A., et al. (2003). *Searching the workplace web.* In the Proceedings of WWW 2003, Budapest, Hungary.

Freeman, E., & Gelernter, D. (1996). Lifestreams: a storage model for personal data. *ACM SIGMOD Record, 25*(1), 80-86.

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (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., & 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.

Gomez, L. M., Lochbaum, C. C., & Landauer, T. K. (1990). All the right words: Finding what you want as a function of richness of indexing vocabulary. *Journal of the American Society for Information Science, 41*(8), 547-559.

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., & Leadem, E. (2003, September 28 - October 3). *Iterative design of metadata creation tools for resource authors.* In the Proceedings of 2003 Dublin Core Conference: Supporting Communities of Discourse and Practice—Metadata Research and Applications, Seattle, Washington.

Grudin, J. (1988). *Why CSCW applications fail: problems in the design and evaluation of organizational interfaces.* In the Proceedings of the ACM conference on Computer-supported cooperative work.

Grudin, J. (2006). *Enterprise Knowledge Management and Emerging Technologies.* In the Proceedings of HCIC 2006 Workshop, Winter Park, CO.

Harter, S. P. (1992). Psychological relevance and information science. *Journal of the American Society for Information Science, 43*(9), 602-615.

Hertzum, M. (1999). *Six Roles of Documents in Professionals' Work.* In the Proceedings of Sixth European Conference on Computer-Supported Cooperative Work, Copenhagen, Denmark.

Hertzum, M., & 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.

Jacob, E. K. (1995). Communication and category structure: the Communicative Process as a Constraint on the Semantic Representation of Information. In B. H. Kwasnik, P. Smith, R. Fidel & C. Beghtol (Eds.), *Advances in Classification Research, Vol. 4*. Medford, NJ: Information Today.

Jian, G., & Jeffres, L. (2006). Understanding Employees' Willingness to Contribute to Shared Electronic Databases: A Three Dimensional Framework. *Communication Research, 33*(4), 242-261.

Jones, W., Phuwanartnurak, A. J., Gill, R., & Bruce, H. (2005). *Don't take my folders away! Organizing personal information to get things done.* In the Proceedings of SIGCHI Conference on Human factors in computing systems, Portland, OR, USA.

Jurafsky, D., & Martin, J. (1999). Word-Sense Disambiguation and Information Retrieval. In *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (pp. 631-666). Place Published: Prentice-Hall.

Kankanhalli, A., Tan, B. C. Y., & Wei, K.-K. (2005). Understanding seeking from electronic knowledge repositories: An empirical study. *Journal of the American Society for Information Science and Technology, 56*(11), 1156 - 1166.

Karau, S. J., & Williams, K. D. (2001). Understanding Individual Motivation in Groups: the Collective Effort Model. In M. E. Turner (Ed.), *Groups at Work: Theory and Research* (pp. 113-141). Mahwah, NJ: Lawrence Erlbaum.

Karger, D. R., Bakshi, K., Huynh, D., Quan, D., & Sinha, V. (2005). *Haystack: A General Purpose Information Management Tool for End Users of Semistructured Data.* In the Proceedings of CIDR 2005.

Kaye, J. J., Vertesi, J., Avery, S., Dafoe, A., David, S., Onaga, L., et al. (2006). *To have and to hold: exploring the personal archive.* In the Proceedings of Proceedings of the SIGCHI conference on Human Factors in computing systems, Montreal, Quebec, Canada.

Kellogg, W. A., & Breen, T. J. (1987). *Evaluating user and system models: applying scaling techniques to problems in human-computer interaction.* In the Proceedings of Proceedings of the SIGCHI conference on Human factors in computing systems and graphics interface, Toronto, Ontario, Canada.

Krauss, R. M., & Fussell, S. R. (1991). Constructing Shared Communicative Environments. In L. B. Resnick, J. M. Levine & S. D. Teasley (Eds.), *Perspectives on Socially Shared Cognition* (pp. 172-201). Washington DC: American Psychological Association.

Kwasnik, B. H. (1991). The Importance of Factors That Are Not Document Attributes in the Organization of Personal Documents. *Journal of Documentation, 47*(4), 389-398.

Lansdale, M. W. (1988). The Psychology of Personal Information Management. *Applied Ergonomics, 19*(1), 55-66.

Leckie, G. J., Pettigrew, K. E., & Sylvain, C. (1996). Modeling the information seeking of professionals: A general model derived from research on engineers, health care professionals, and lawyers. *Library Quarterly, 66*(2), 161-193.

Levin, D. Z., & Cross, R. (2004). The strength of weak ties you can trust: the mediating role of trust in effective knowledge transfer. *Management Science, 50*(11), 1477-1490.

Lewis, D. D., & Jones, K. S. (1996). Natural language processing for information retrieval. *Commun. ACM, 39*(1), 92-101.

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.

Marchionini, G. (1997). Foundations for Personal Information Infrastructures: Information-Seeking Knowledge, Skills, and Attitudes. In *Information Seeking in Electronic Environments*: Cambridge University Press.

Mark, G. (1997, 19 - 28). *Merging multiple perspectives in groupware use: intra- and intergroup conventions.* In the Proceedings of the international ACM SIGGROUP conference on Supporting group work: the integration challenge.

Mark, G., & Prinz, W. (1997). *What Happened to our Document in the Shared Workspace? The Need for Groupware Conventions.* In the Proceedings of IFIP TC13 International Conference on Human-Computer Interaction.

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., & Davis, M. (2006). *Position Paper, Tagging, Taxonomy, Flickr, Article, ToRead.* In the Proceedings of WWW 2006 Collaborative Web Tagging Workshop, Edinburgh, Scotland.

McCreadie, M., & Rice, R. E. (1999). Trends in analyzing access to information. Part I: cross-disciplinary conceptualizations of access. *Information Processing & Management, 35*(1), 45-76.

Mervis, C. B., & Rosch, E. (1981). Categorization of Natural Objects. *Annual Review of Psychology, 32*, 89-115.

Murphy, G. L., & Lassaline, M. E. (1997). Hierarchical Structure in Concepts and the Basic Level of Categorization. In K. Lamberts & D. Shanks (Eds.), *Knowledge, Concepts, and Categories (Studies in Cognition)* (pp. 93-131). Hove, East Sussex, UK: Psychology Press.

Olson, J. S., & Olson, G. M. (1999). Computer supported cooperative work. In F. T. Durso (Ed.), *Handbook of applied cognition* (pp. 409-442). Chichester, England: John Wiley & Sons.

Palen, L., & Grudin, J. (2002). Discretionary adoption of group support softwware. In B. E. Munkvold (Ed.), *Organizational implementation of collaboration technology* (pp. 159-180): Springer Verlag.

Pirolli, P. (2005). Rational Analyses of Information Foraging on the Web. *Cognitive Science: A Multidisciplinary Journal, 29*(3), 343-373.

Prinz, W., & Zaman, B. (2005). *Proactive support for the organization of shared workspaces using activity patterns and content analysis.* In the Proceedings of 2005 international ACM SIGGROUP conference on Supporting group work, Sanibel Island, Florida, USA.

Quan, D., Bakshi, K., Huynh, D., & Karger, D. R. (2003). *User Interfaces for Supporting Multiple Categorization.* In the Proceedings of INTERACT 2003.

Reitman, J. S., & Rueter, H. H. (1980). Organization revealed by recall orders and confirmed by pauses. *Cognitive Psychology, 12*(4), 554-581.

Ringel, M., Cutrell, E., Dumais, S. T., & Horvitz, E. (2003, September 2003). *Milestones in time: The value of landmarks in retrieving information from personal stores.* In the Proceedings of Interact 2003.

Russell, A. W., & Schober, M. F. (1999). How beliefs about a partner's goals affect referring in goal-discrepant conversations. *Discourse Processes, 27*(1), 1-33.

Šauperl, A. (2004). Catalogers' common ground and shared knowledge. *Journal of the American Society for Information Science and Technology, 55*(1), 55-63.

Schamber, L. (1994). Relevance and Information Behavior. *Annual Review of Information Science and Technology (ARIST), 29*, 3-48.

Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology, 21*(2), 211-232.

Star, S. L., & Ruhleder, K. (1996). Steps Towards an Ecology of Infrastructure: Design and Access for Large-Scale Systems. *Information Systems Research*(7), 111-138.

Suchman, L. (1994). Do Categories Have Politics? The Language/Action Perspective Reconsidered. *Computer Supported Cooperative Work, 2*(3), 177-190.

Suthers, D., & Hundhausen, C. (2002, January 7-11, 2002). *The Effects of Representation on Students' Elaborations in Collaborative Inquiry.* In the Proceedings of CSCL 2002, Boulder, Colorado.

Teevan, J., Alvarado, C., Ackerman, M. S., & Karger, D. R. (2004). *The perfect search engine is not enough: a study of orienteering behavior in directed search.* In the Proceedings of SIGCHI conference on Human factors in computing systems, Vienna, Austria.

Teevan, J., Dumais, S. T., & Horvitz, E. (2005). *Personalizing Search via Automated Analysis of Interests and Activities.* In the Proceedings of SIGIR'05, Salvador, Brazil.

Trigg, R. H., Blomberg, J., & Suchman, L. (1999). *Moving document collections online: The evolution of a shared repository.* In the Proceedings of Sixth European Conference on Computer-Supported Cooperative Work, Copenhagen, Denmark.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*(2), 207-232.

Voida, S., Edwards, W. K., Newman, M. W., Grinter, R. E., & Ducheneaut, N. (2006). *Share and share alike: exploring the user interface affordances of file sharing.* In the Proceedings of Proceedings of the SIGCHI conference on Human Factors in computing systems, Montreal, Quebec, Canada.

Wageman, R. (2001). The Meaning of Interdependence. In M. E. Turner (Ed.), *Groups at Work: Theory and Research* (pp. 197-217). Mahwah, NJ: Lawrence Erlbaum.

Whittaker, S., & 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., & Sidner, C. (1996). *Email overload: exploring personal information management of email.* In the Proceedings of CHI '96: Human factors in computing systems, Vancouver, British Columbia.

Winograd, T. (1994). Categories, Disciplines, and Social Coordination. *Computer Supported Cooperative Work, 2*(3), 191-197.

Woods, D. D. (1998). Designs are hypotheses about how artifacts shape cognition and collaboration. *Ergonomics, 41*(2), 168-173.

Wright, D. B., Mathews, S. A., & Skagerberg, E. M. (2005). Social Recognition Memory: The Effect of Other People's Responses for Previously Seen and Unseen Items. *Journal of Experimental Psychology: Applied, 11*(3), 200-209.

Zhang, J., & Norman, D. A. (1994). Representations in distributed cognitive tasks. *Cognitive Science, 18*(1), 87-122.

Appendix A: List of Research Questions

Part I: Storing

1.1: What are the effects of interdependence and ownership norms on information producers' attitudes toward sharing, and contributions to a shared file repository?

1.2: What factors lead to the successful adoption of shared file repositories by information producers?

1.3: What is an optimal ratio of information producers to consumers to sustain continued use of a shared file repository? How is this affected by the purpose for which the repository is used, and the type of content stored?

Part II: Organizing

2.1: How do file naming and labeling conventions evolve in shared file repositories? How are they enforced, or reinforced?

2.2: What evidence for a "shared basis" or "feeling of others' knowing" exists and can be communicated in a shared file repository?

2.3: How does common ground affect information producers' choices of labels for files?

2.4: To what extent do information producers "package" their contributions to a shared file repository? How does this affect repository use by information consumers?

2.5: What are the goals of information producers and consumers, related to their use of shared file repositories?

2.6: How do the strategies and tactics of information producers affect the structure of the shared file repository, and the strategies and tactics of information consumers?

Part III: Seeking

3.1: For what purposes are shared file repositories used? What are information producers' and consumers' goals? What files in a repository are used most often, who uses them, and why?

3.2: Does the strength of the ties in the social network of shared file repository users predict the level of participation by information consumers?

Part IV: Finding

4.1: Is tagging more like labeling or assigning something to a category? What are the implications of this for shared file repositories?

4.2: How does the amount of effort expended to select "good" tags affect search and retrieval of tagged items?

4.3: Are social tagging systems a solution to the vocabulary problem for search? How many users are necessary for this solution to be viable? What are the implications of common ground for social tagging systems?

4.4: How do different external representations (full-text search, metadata+full-text, or navigating/orienteering a hierarchy) affect the structure of users' internal representations for a shared file repository?

4.5: Do basic-level categories appear in personal or shared file repositories? What is their effect on how the repositories are used?

4.6: How do information consumers' internal representations for a shared file repository correspond with the external representation?

4.7: How do aspects of a file's history or context interact with the external representation format of the repository to affect an information consumer's relevance judgments and search process?

4.8: What factors influence the performance of different external representations (i.e. full-text search, metadata plus full-text, or navigating/orienteering a hierarchy) for finding information in a shared file repository? How does this depend upon the purpose and context within which the repository is used?

Just email it to me!