

# What did you call it again? Language Use in Group Information Management Systems

Emilee Rader

October 1, 2007

## Contents

1	Introduction . . . . .	1
1.1	Shared File Repositories . . . . .	2
1.2	Language Use . . . . .	3
1.3	Proposed Research . . . . .	4
2	Literature Review . . . . .	4
2.1	File Naming Conventions . . . . .	4
2.2	Vocabulary Problem and Common Ground . . . . .	5
2.3	File Location . . . . .	7
2.4	Information Seeking in Shared File Repositories . . . . .	7
2.5	Navigation vs. Search . . . . .	9
2.6	Concept Map . . . . .	10
3	Research Plan . . . . .	10
3.1	Study 1: CTools Project Sites . . . . .	10
3.2	Study 2A: Naming and Organizing . . . . .	14
3.3	Study 2B: Information Seeking . . . . .	16
3.4	Significance and Impact . . . . .	17
4	Timeline and Proposed Budget . . . . .	18
4.1	Timeline . . . . .	18
4.2	Budget . . . . .	18

## 1 Introduction

Much of an organization’s information is represented in the form of documents, such as reports, memos, meeting minutes, email messages, etc. Ineffective document management incurs costs such as “lost work time, ineffective access to information, duplication of effort, failure to share information, and information overload” (Gordon, 1997). Many different kinds of workgroups including research labs, corporate teams, and software developers use central online repositories for storing information. Some examples include enterprise content management systems, software code repositories, and wikis or blogs used by teams as a form of “group memory.”

Online repositories are maintained by many organizations. They are essential for document sharing, and can be greatly beneficial for organizational efficiency, communicating organizational goals, and also for learning and innovation. They can contain “mission critical informa-

tion” such that if it were lost there would be serious consequences (Blair, 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. Instead, they tend to accumulate content over time and become more and more disorganized, such that users have difficulty finding the files they need. According to (Gordon, 1997):

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

The research proposed in this document examines situations in which users are not able to effectively find and access information in shared file repositories, and suggests remedies for these problems.

## 1.1 Shared File Repositories

Shared file repositories are online storage spaces used by workgroups for storing, organizing, and sharing documents and other files, and their use is increasing. A shared file repository is more complex than just an “aggregate of every individual’s contribution” (Jian and Jeffres, 2006), and maintaining it is a collaborative activity. A repository user is generally familiar with his fellow group members, and with projects and joint work activities they are engaged in together. However, he can expect to be familiar with only some of the files stored in a shared file repository, and he 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. This can be frustrating enough that users seeking information circumvent shared file repositories altogether, opting sometimes to request files from others via email instead.

Several aspects of shared file repositories set them apart from other ways information is stored and shared online. Some users are information *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 that can exist 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. People differ in the ways in which they prefer to structure their personal file repositories, which can be a problem for information seeking in a shared repository (Berlin et al., 1993; Whittaker and Hirschberg, 2001). In addition, unlike libraries or even an organization’s website designed by an information architect, shared file repositories generally do not have a structured organization scheme.

Library records and classification schemes were created for describing items and codifying relationships between subjects (Rafferty, 2001). In contrast, shared file repositories do not necessarily have unified goals or purposes to guide users’ choices. Being *unstructured* means

that less effort is required when storing and labeling files. Because there are no rules describing suitable contributions users are free to express what is important to them about the content, rather than what might fit within the classification scheme, within the constraints of what the system will allow (Marlow et al., 2006). People are also more willing to voluntarily contribute metadata when the task is less onerous (Greenberg et al., 2003). However, the lack of structure can also have a down side. The act of storing and labeling files in a shared file repository is one of packaging content for later reuse, and in most situations, people do not effectively package content for others (Markus, 2001). Effective packaging requires that one consider the information needs and context of the reusers, or information consumers. Information producers' unstructured choices can affect the future reuse of the information in the repository by information consumers, because decisions about file names and locations provide the means by which the user-contributed content may be found and accessed. The distinction between those producing the content and those consuming it is an important one: it implies that there is communication going on between users who are producing content, and users who are accessing content, via the repository. And because file and folder hierarchies are represented in the form of text, factors that shape language choices in communication situations might also affect packaging choices for user-contributed content.

## 1.2 Language Use

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

There is much experimental evidence to support the idea that common ground affects language use. Speakers tailor their utterances for listeners, with performance implications (Schober and Clark, 1989). In addition, people create labels for their own use that are different from those created for an unknown future person (Fussell and Krauss, 1989). People tailor what they say to whomever is the intended recipient; it is reasonable to think that common ground might indeed affect the names information producers create for files they store in a shared file repository. An important difference between a real-time conversation and any type of asynchronous communication is in the timing of feedback, which is important for establishing common ground and negotiating meaning (Clark and Brennan, 1991). For example, facial expressions are a form of nonverbal feedback that convey whether or not a speaker has been understood by a listener. Systems allowing for the provision of user-contributed content rarely include a mechanism for feedback; indeed, what information might be included in that feedback, and what form it might take, are open questions.

### 1.3 Proposed Research

The purpose of this research is to improve users' ability to find information in situations where multiple users access shared information resources via a user-contributed content system, by investigating how language use affects the organization, structure and seeking of information. In my dissertation, I ask the following questions:

How does language use affect information sharing in user-contributed content repositories?

1. How do groups organize and share information via an online repository of user-contributed content? What problems do individual members encounter during information seeking within the repository?
2. How do the source of common ground and the intended audience affect file and folder name choices, and the effectiveness of those names for information seeking?

I will first conduct a field study a user-contributed content system at the University of Michigan (tools.umich.edu). I completed pilot interviews with faculty, staff and students using CTools to support ongoing collaborative projects. I will follow up the pilot with a more focused investigation of information seeking by group members within their own shared file repositories. I am also planning a series of experiments subsequent to the field study that will test hypotheses about the effects of common ground on choices made by users when storing, organizing, and seeking files using shared file repositories.

## 2 Literature Review

### 2.1 File Naming Conventions

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

Sometimes, conventions are agreed to in principle, and then intentionally ignored in practice. Mark and Prinz (1997) conducted a field study of a group using a "large groupware system" to store and share 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 and Prinz (1997) concluded that in this case, it had been too difficult to imagine in advance what conventions would be needed. As the system was used, work practices changed, making the conventions the group had agreed to less appropriate for the situations that arose. Also, there were some users who were unwilling to give up their own, idiosyncratic practices. In some cases it was a conscious choice to violate conventions. One user said, "Naming conventions, reference code, and subject area, I always violate. I give file names that seem to fit" ((Mark and Prinz, 1997); p. 23).

At the outset of using a repository, users don't know what information structures will work best, and after it has been in use for a while cleaning up the repository is too onerous a task for

most users to be willing to undertake (Barreau, 1995). Shared 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.

## 2.2 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.

Humans' use of language is imprecise and flexible, and meaning is determined by the surrounding context, and complex communication processes. As a conversation progresses, participants introduce ideas and vocabulary that become part of their common ground, and can subsequently be referred to without the overhead of having to re-introduce them. Common ground is necessary for coordination of conversation, and essential for people to understand one another. Conversation participants believe common ground exists when there is evidence for a "shared basis". Evidence that a shared basis exists for members of a workgroup using a shared 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. Common ground can be classified into three types (Nickerson, 1999):

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

There is much experimental evidence to support the idea that common ground affects language use. Speakers tailor their utterances for listeners, with performance implications. In

an experiment conducted by Schober and Clark (1989), participants completed a referential communication task where one participant instructed another how to construct an abstract shape using puzzle pieces. A third participant (the 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 and Schober (1999) found that being correctly informed about a partner's goals had an impact on how much was said and how understanding was displayed. Also, participants assumed others' goals were the same as theirs if they were not told otherwise as part of the experiment.

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

The results of these experiments indicate that common ground might indeed affect the 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". While a shared file repository is not a communications system, language is being used as abbreviations to represent the contents of files, and also to suggest relationships among groups of files. Common ground helps us understand each other in conversation; the same might be true when the communication is mediated by a shared file repository. Groups with more common ground might label files with names that others in the group will be able to anticipate more frequently than 20% of the time. However, this is not as straightforward as it sounds. (Hertzum and Pejtersen, 2000) wrote:

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

In other words, simply being aware of others' knowledge, background and joint experiences is insufficient for properly "packaging" information for a shared file repository. The ability to take the perspective of others is also necessary. 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 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.

### 2.3 File Location

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 folder location in the hierarchy at which someone else will be able to access the file again. Whittaker and 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 et al., 2005; Malone, 1983; Whittaker and Hirschberg, 2001). The more folders one has, the less helpful they are at reducing the amount of stuff one has to remember. If each folder contains two or three items, there are a lot more folders to remember than if each one contains ten or fifteen items.

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 and Hirschberg, 2001). People feel like they should hang onto information they aren’t sure they need, just in case the need might arise later. Often, later never comes, and people generally don’t go back and purge without an incentive or triggering event. Deferring evaluation might mean information producers never get around to thinking about whether a new file should be stored in the repository or not, meaning the files might not end up in the repository at all.

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

### 2.4 Information Seeking in Shared File Repositories

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 may seem somewhat misleading or incorrect. This is not likely to be intentional, but as I have shown in previous sections, lack of adherence to conventions, unique individual goals and strategies, and the vocabulary problem make it difficult for information consumers to find what they need in a shared file repository. Markus (2001) wrote that information producers tend not to be very good at documenting their work. But she also argued that even when people do a great job at documenting, work “byproducts” like notes and meetings and diagrams etc. can build up to such an extent that too much effort is required to search them:

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

This problem is compounded when the producers and the intended users of the information are not the same people. When information producers document for themselves, they are the beneficiaries of all their hard work. There are few inherent 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 et al., 2003).

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.

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

Once an information consumer has decided that useful information is likely to be present in the repository, he must 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 and 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.

## 2.5 Navigation 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. 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 search interface only. According to Dourish (2004), “In information work, the meaningfulness of information for people’s work is often encoded in the structures by which that information is organized” (p. 30). Jones et al. (2005) found that folder hierarchies and 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 et al., 2006).

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 study by Mobrand and Spyridakis (2007) demonstrated that “navigational link phrasing” — link naming — affected navigation in a news website; confusing or ambiguous hyperlinks decreased overall comprehension of the information, and discouraged exploration. An experiment conducted by Vaughan and Dillon (2006) found similar results; they created two versions of the same health information website, one which was similar in design and link naming to typical health information websites, and one which violated users’ expectations for page layout and link naming. The group that used the “expectation-conforming” version of the website explores more of the site initially than the group using the “expectation-violating” version, but when asked to search for information they were able to find what they were looking for faster.

A shared 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 et al., 2004; Trigg et al., 1999). In a study conducted by Boardman and Sasse (2004) users searching their personal repositories used a combination of browsing and sorting of folders. Because they were searching their 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 et al. (2004) called this *orienteering*: using recall to make an initial jump to a location from which to start navigating in steps, via recognition, toward the ultimate goal. At each stage, the local context is used to remind people about where they should go for the next step. Teevan et al. (2004) mentioned one participant who tried to find something in her personal 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.

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

## 2.6 Concept Map

The concept map shown in Figure 1 on page 11 illustrates the relationships among various factors hypothesized to affect file naming and organizing in group information management situations, and subsequent information seeking. Three research studies are proposed in this document to investigate these relationships.

The first study is a field study of information seeking behavior, designed to explore the relationship between the existing organization and naming structure of a shared file repository and its utility for information seeking. The field study will also collect general information about how this type of group information management system is used in a real-world setting, because the research literature is fairly sparse in this respect. A series of experiments will focus on common ground and awareness of potential future reuse, and their impact on file-names and information organizational structures. The mixed-methods approach is important for this research project, because a lab experiment is limited in the degree of external validity it can achieve. Studying the behaviors in question in the field is an essential complement to the laboratory research.

- Study 1 (field study): How do groups organize and share information via an online repository of user-contributed content? What problems do individual members encounter during information seeking within the repository?
- Study 2A (experiment): How do the type of common ground (from community membership or shared past experience) and the intended audience affect file and folder name choices?
- Study 2B (experiment): How does the influence of common ground and intended audience on file and folder names affect information seeking outcomes in a shared online repository?

## 3 Research Plan

### 3.1 Study 1: CTools Project Sites

CTools project sites are sites created not for courses, but to support group projects that take place at the University of Michigan. Anyone may create a project site. CTools project sites are primarily used for sharing files via the “Resources” section of the site, which presents users

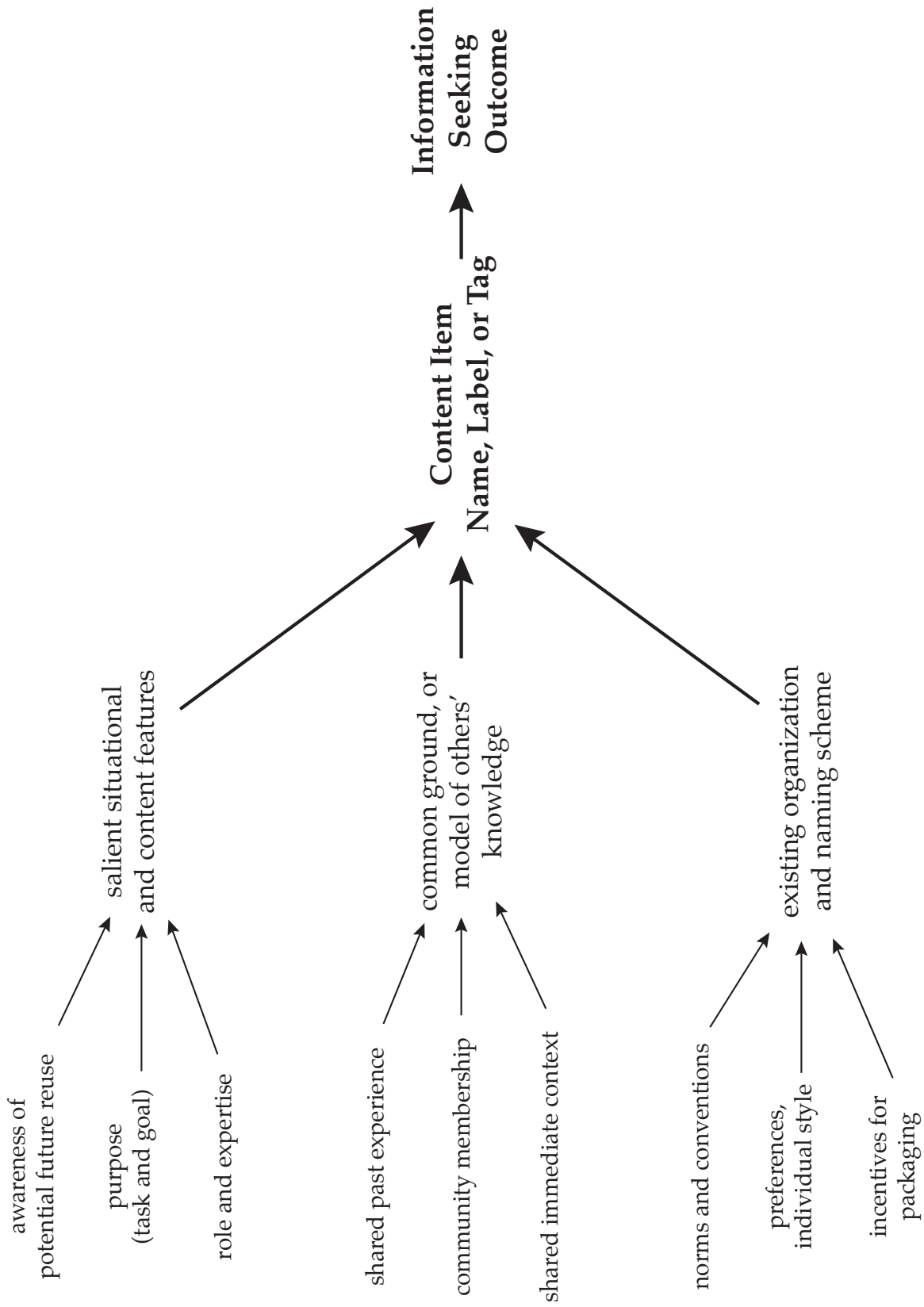


Fig. 1: Concept map depicting hypothesized influences on naming choices and information seeking in user-contributed content systems

with an interface that follows the desktop file-and-folder metaphor for organizing information. I propose conducting a study of information seeking with users of CTools project sites. CTools is an example of a group information management system in use by thousands of people, and it is possible to study their use of the system in the environment in which it occurs. Results will allow me to quantify and describe information seeking activities in greater detail, in the context of the particular system. It is important to investigate people using their own content; artificial information seeking tasks in the lab are less externally valid.

### **Research Question**

How do groups organize and share information via an online repository of user-contributed content? What problems do individual members encounter during information seeking within the repository?

### **Sources of Data: Interviews, Search Tasks and Event Logs**

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

- frequency and recency of use
- ownership of the item
- depth in the hierarchy
- descriptiveness of filename, in relation to the information contained in the file

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

### **Informant Interview Questions**

- Tell me about the site. Whats the purpose? Who is the group? Whats your role? How long has it been around?
- Tell me about how you use the site... how often, for what kinds of things? What kinds of items are in the site? Who are the other primary users?
- Can you give me a tour of the site – talk about each folder and what is in it?

- When was the last time you used the repository before this study session? Tell me about that time – what were you doing – why did you need the file? Can you find that file for me now (think out loud)? When was the time before that?
- Can you think of a file that is an important file, or used often by the group? Show me. Open the file, tell me about it... Another file? What parts of the site get a lot of use? Not very much use?
- When was the last time you added a file? Tell me about it... can you remember how you went about deciding what to name it and where to put it? Is there something you've been meaning to add? Show me...
- When was the last time you renamed, moved, or deleted something? etc.
- When was the last time someone in your group cleaned up the site? Tell me about it... what happened?

#### Other Site Member Interview Questions

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

#### CTools Event Logs

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

#### Limitations and Tradeoffs

Because I am planning to collect data with respect to one particular system (CTools), this research should be considered a case study. Each system has a different interface design that supports user needs and goals to varying degrees of success; given unlimited time and

resources, it would be better to use the same interview and search task protocol with participants using different group information management software.

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

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

### **Participants**

I plan to recruit participants from 8-10 different project sites. I am hoping to get a range of different site types but this will depend on recruitment and who is willing to participate. I am looking for sites that have been around for at least 6 months, and that have at least 5 members, one of whom will serve as the informant.

### **Measures and Analysis**

Audio and users' interactions with the CTools site will be recorded for analysis, using Tech-Smith's Morae software [website]. Digital audio will also be recorded throughout the session. Think aloud protocols from the search tasks will be analyzed, and qualitative coding analysis techniques will be used on the interview answers. Quantitative analyses will consist of counts of search task successes and failures, as well as distance measures of participants' initial guesses as to where the target items might be located in the hierarchy.

## **3.2 Study 2A: Naming and Organizing**

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

### **Research Question**

How does type of common ground (from community membership or shared past experience) and the intended audience affect file and folder name choices?

## Participants

Participants will be Master's students in the School of Information (MSI) who took the course SI 501 in either Fall 2006 (MSI Class of '08) or Fall 2007 (MSI Class of '09), and non-SI graduate students. Common ground in this experiment is a subject variable; I am not actually manipulating it. Students who took SI 501 at the same time can be said to have common ground from *shared past experience*, students from SI in different cohorts have *community membership* common ground. The information they will be asked to name and organize will be related to topics in SI 501.

## Method

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

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

The levels of *intended audience* are:

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

Participants will be given a set of documents (related in subject matter to material covered in SI 501) to name and organize into folders. They will be provided with instructions regarding their intended audience; i.e., who will be using the organization hierarchies they create. See Figure 2 on page 16 for the breakdown of conditions and number of participants in each condition.

After organizing the documents, participants will be asked to complete a post-questionnaire asking them about the choices they made during the task. For example, the questionnaire will be used to find out whether participants report considering the information needs or expectations of their intended audience when creating the names and organization hierarchy. Finally, the questionnaire will also include asking participants to describe their intended audience, so that it might be possible to get an idea of the model of others' knowledge upon which their choices were based.

## Measures and Analysis

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

	No Audience	For Self	For Your Class	For Last Year's Class	For General Public	(total)
<b>MSI Class of 2008</b>	control 10	control 10	SPE 10	CM 10	NCG 10	(50)
<b>MSI Class of 2009</b>	control 10	control 10	SPE 10	CM 10	NCG 10	(50)
<b>Non-SI Grad Student</b>	control 10	control 10	X (no subjects)	X (no subjects)	NCG 10	(22)
(total)	(30)	(30)	(20)	(20)	(30)	(130)

SPE = Shared Past Experience Common Ground  
 CM = Community Membership Common Ground  
 NCG = No Common Ground

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

questionnaire data will be analyzed, and open-ended questions will be analyzed using qualitative coding techniques.

## Hypotheses

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

## 3.3 Study 2B: Information Seeking

In this second experiment, participants will return and complete search tasks using the information organization hierarchies created in the first experiment. This experiment is designed to investigate the impact of information organization hierarchies created by others with whom different kinds of common ground are shared and that were created with different audiences in mind, on information seeking outcomes.

### Research Question

How does the influence of common ground and intended audience on file and folder names affect information seeking outcomes in a shared online repository?

## Participants

At the end of the first experiment, appointments will be set up with participants to return 4-6 weeks later for the second experiment.

## Method

In this experiment, participants will return to the lab and complete information seeking tasks using the hierarchies that were created in the first experiment. Search targets will be a subset of the documents they organized several weeks earlier. The experiment will be a 3 (cohort, between) x 6 (hierarchy type, within) mixed design. So, for example, a minimum of 20 participants returning from each cohort will be asked to complete search tasks using the following hierarchies (see Figure 2 on page 16 for the table laying out all the conditions):

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

## Measures and Analysis

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

## Hypotheses

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

## 3.4 Significance and Impact

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

## 4 Timeline and Proposed Budget

### 4.1 Timeline

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

January 2008: Participant recruiting

February – April 2008: Data collection

Summer 2008: Data analysis

September – December 2008: Writing

January – February 2009: Finished!

### 4.2 Budget

Expenses for this research consist primarily of subject payments. I plan to use Morae to capture detailed task timings in the experiments, but will need to either purchase a license or secure access to computing equipment owned by SI in order to use software licensed or purchased by the university. The cost of a single copy of Morae Recorder is \$195. Subject payments will be around \$2800. I am planning to apply for a Rackham Discretionary Funds grant to support this work.

*Study 1: CTools Project Sites* \$900. 8-10 different sites, 1 informant (\$25) and 4 users per site (\$15). Upper bound approx \$850. (plus one pilot site from SI if possible, 1 informant and 2 users = \$55).

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

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

## References

- Ackerman, M. S. (2000). The intellectual challenge of cscw: The gap between social requirements and technical feasibility. *Human-Computer Interaction*, 15(2):181 – 203.
- Barreau, D. (1995). Context as a factor in personal information management systems. *Journal of the American Society for Information Science*, 46(5):327–339.
- Bates, M. J. (1998). Indexing and access for digital libraries and the internet: Human, database, and domain factors. *Journal of the American Society for Information Science*, 49(13):1185–1205.
- Bellotti, V., Ducheneaut, N., Howard, M., Smith, I., and Grinter, R. (2005). Quality vs. quantity: Email-centric task-management and its relationship with overload. *Human-Computer Interaction*, 20(1-2):89–138.
- Berlin, L. M., Jeffries, R., O'Day, V. L., Paepcke, A., and Wharton, C. (1993). Where did you put it? issues in the design and use of a group memory. In *SIGCHI conference on Human factors in computing systems*, pages 23–30, Amsterdam, The Netherlands. ACM Press.

- Blair, D. C. (2002). Information retrieval and the philosophy of language. In Cronin, B., editor, *Annual Review of Information Science and Technology*, volume 37, pages 3–50. The American Society for Information Science and Technology, Medford, NJ.
- Boardman, R. and Sasse, M. A. (2004). "stuff goes into the computer and doesn't come out": A cross-tool study of personal information management. In *SIGCHI Conference on Human Factors in Computing Systems*, pages 583–590, Vienna, Austria.
- Bruce, H., Jones, W., and Dumais, S. (2004). Information behaviour that keeps found things found. *Information Research*, 10(1).
- Chalmers, M. (2003). Informatics, architecture and language. In Hook, K., Benyon, D., and Munro, A. J., editors, *Designing Information Spaces: The Social Navigation Approach*, pages 315–342. Springer, London.
- Clark, H. H. (1996). Common ground. In *Using Language*. Cambridge University Press, Cambridge.
- Clark, H. H. and Brennan, S. E. (1991). Grounding in communication. In Resnick, L. B., Levine, J. M., and Teasley, S. D., editors, *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association, Washington DC.
- Cutrell, E., Robbins, D., Dumais, S., and Sarin, R. (2006). Fast, flexible filtering with phlat. In *CHI '06*, pages 261–270, Montreal, Quebec, Canada. ACM Press.
- Dourish, P. (2004). What we talk about when we talk about context. *Personal and Ubiquitous Computing*, 8(1):19–30.
- Furnas, G., Landauer, T., Gomez, L., and Dumais, S. (1983). Statistical semantics: Analysis of the potential performance of key-word information systems. *The Bell System Technical Journal*, 62(6):1753–1806.
- Fussell, S. R. and Krauss, R. M. (1989). The effects of intended audience on message production and comprehension: Reference in a common ground framework. *Journal of Experimental Social Psychology*, 25(3):203–219.
- Gordon, M. D. (1997). It's 10 a.m. do you know where your documents are? the nature and scope of information retrieval problems in business. *Information Processing & Management*, 33(1):107–122.
- Greenberg, J., Crystal, A., Robertson, W. D., and Leadem, E. (2003). Iterative design of metadata creation tools for resource authors. In *2003 Dublin Core Conference: Supporting Communities of Discourse and Practice—Metadata Research and Applications*, Seattle, Washington.
- Hertzum, M. (1999). Six roles of documents in professionals' work. In *Sixth European Conference on Computer-Supported Cooperative Work*, Copenhagen, Denmark.
- Hertzum, M. and Pejtersen, A. M. (2000). The information-seeking practices of engineers: searching for documents as well as for people. *Information Processing & Management*, 36(1):761–778.
- Jian, G. and Jeffres, L. (2006). Understanding employees' willingness to contribute to shared electronic databases: A three dimensional framework. *Communication Research*, 33(4):242–261.
- Jones, W., Phuwanartnurak, A. J., Gill, R., and Bruce, H. (2005). Don't take my folders away!

- organizing personal information to get things done. In *SIGCHI Conference on Human factors in computing systems*, pages 1505–1508, Portland, OR, USA. ACM Press.
- Krauss, R. and Fussel, S. (1991). Perspective-taking in communication: representations of others' knowledge in reference. *Social Cognition*, 9(2-24).
- Lansdale, M. (1988). The psychology of personal information management. *Applied Ergonomics*, 19(1):55–66.
- Malone, T. W. (1983). How do people organize their desks? implications for the design of office information systems. *ACM Transactions on Information Systems (TOIS)*, 1(1):99 – 112.
- Mark, G. and Prinz, W. (1997). What happened to our document in the shared workspace? the need for groupware conventions. In *IFIP TC13 International Conference on Human-Computer Interaction*, pages 413–420.
- Markus, L. M. (2001). Toward a theory of knowledge reuse: Types of knowledge reuse situations and factors in reuse success. *Journal of Management Information Systems*, 18(1):57 – 93.
- Marlow, C., Naaman, M., boyd, d., and Davis, M. (2006). Position paper, tagging, taxonomy, flickr, article, toread. In *WWW 2006 Collaborative Web Tagging Workshop*, Edinburgh, Scotland.
- Mobrand, K. A. and Spyridakis, J. H. (2007). Explicitness of local navigational links: comprehension, perceptions of use, and browsing behavior. *Journal of Information Science*, 33(1):41–61.
- Nickerson, R. S. (1999). How we know – and sometimes misjudge – what others know: imputing one's own knowledge to others. *Psychological Bulletin*, 125(6):737–759.
- Pirolli, P. (2005). Rational analyses of information foraging on the web. *Cognitive Science*, 29(3):343–373.
- Rafferty, P. (2001). The representation of knowledge in library classification schemes. *Knowledge Organization*, 28(4):180–191.
- Russell, A. W. and Schober, M. F. (1999). How beliefs about a partner's goals affect referring in goal-discrepant conversations. *Discourse Processes*, 27(1):1–33.
- Schober, M. F. and Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, 21(2):211–232.
- Suchman, L. (1994). Do categories have politics? the language/action perspective reconsidered. *Computer Supported Cooperative Work*, 2(3):177–190.
- Teevan, J., Alvarado, C., Ackerman, M. S., and Karger, D. R. (2004). The perfect search engine is not enough: a study of orienteering behavior in directed search. In *SIGCHI conference on Human factors in computing systems*, pages 415–422, Vienna, Austria. ACM Press.
- Trigg, R. H., Blomberg, J., and Suchman, L. (1999). Moving document collections online: The evolution of a shared repository. In *Sixth European Conference on Computer-Supported Cooperative Work*, Copenhagen, Denmark.
- Šaupperl, A. (2004). Catalogers' common ground and shared knowledge. *Journal of the American Society for Information Science and Technology*, 55(1):55–63.

- Vaughan, M. W. and Dillon, A. (2006). Why structure and genre matter for users of digital information: A longitudinal experiment with readers of a web-based newspaper. *International Journal of Human-Computer Studies*, 64(6):502–526.
- Whittaker, S. and Hirschberg, J. (2001). The character, value, and management of personal paper archives. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 8(2):150 – 170.
- Whittaker, S. and Sidner, C. (1996). Email overload: exploring personal information management of email. In *CHI '96: Human factors in computing systems*, pages 276–283, Vancouver, British Columbia.