How Knowledge Workers Gather Information from the Web: Implications for Peer-to-Peer File Sharing Tools[1]

Jennifer Hyams, Abigail Sellen
Mobile and Media Systems Laboratory
HP Laboratories Bristol
HPL-2003-95
May 19th , 2003*

E-mail: jen_hyams@yahoo.co.uk, abigailsellen@hp.com

peer-to-peer, knowledge workers, information lifecycle, information sharing, web use, information gathering

The success of peer-to-peer (p2p) music-sharing has no doubt contributed to assumptions that individuals' PCs are a vast untapped resource of assets just waiting to be unlocked. This includes the push for opening up our file spaces at work to allow peers access to previously inaccessible information. We explore the potential of these ideas and test some of the assumptions underlying them by looking at 16 knowledge workers' file spaces in the context of Web information-gathering tasks. Knowledge workers' file spaces are more like "workbenches" than "archives" and the information held within them is fundamentally different to that which is placed in shared information spaces. Work is carried out on information to make it shareable, yet this information is found side-by-side on the "workbench" with unshareable information. This leads us to question the potential value of enabling people to open up their file spaces without considering the reusability of this information for others.

* Internal Accession Date Only Approved for External Publication
[1] People and Computers XVII, Proceedings of HCI 2003 : Designing for Society, September 2003, Bath, UK

How Knowledge Workers Gather Information from the Web: Implications for Peer-to-Peer File Sharing Tools

Jennifer Hyams and Abigail Sellen

Hewlett-Packard Labs, Filton Road, Stoke Gifford,
Bristol. BS34 8QZ. UK
jen_hyams@yahoo.co.uk; abigailsellen@hp.com

The success of peer-to-peer (p2p) music sharing has no doubt contributed to assumptions that individuals' PCs are a vast untapped resource of assets just waiting to be unlocked by such systems. This includes the push for opening up our file spaces at work to allow peers access to previously inaccessible information with minimum effort. We wished to explore the potential value of these ideas and to test some of the assumptions underlying them, the motivation being that we believed the issues raised by this investigation would be important to those developing p2p information sharing tools. We do this by looking at the flow of information in and out of 16 knowledge workers' file spaces in the context of carrying out Web information gathering tasks at work. In doing this we find that the file spaces used for knowledge work are more like "workbenches" than "archives" and that the information held within them is fundamentally different in content and organisation to that which knowledge workers place in shared information spaces such as the Web. Knowledge workers work on their information to make it shareable to specific audiences yet this information is found side by side on the "workbench" with unshareable information. This leads us to question the potential value of enabling people to open up their file spaces without having regard to the reusability of this information for others.

Keywords: Peer-to-peer, Knowledge workers, Information lifecycle, Information sharing, Web use, Information gathering.

1 Introduction

The rise and subsequent fall of the music file sharing site, Napster, not only created great furore in the music industry and new dilemmas for copyright law, it also raised awareness of the potential popularity for new kinds of tools and applications which work in a decentralised way. These new models of computing, known more generally as "peer-to-peer" (p2p) computing, hold out the promise of opening up previously unused or inaccessible resources from the "vast untapped resource of personal computers owned by ordinary people" [Kubiatowicz 2003].

While there is some dispute over the proper definition of p2p architectures, this vision is one in which the role of server-based networks is either minimised or bypassed altogether, allowing people to directly share resources (be they storage, cycles or content) between people, or more accurately between people's individual PC's. These concepts take different forms. For example, grid computing describes the ability to share processing power and storage capacity across institutional borders and across clusters of individual computers. Other concepts are more clearly directed at the ability to share multimedia files, bookmarks, educational materials, work-based documents or other kinds of information, usually within specific communities or groups [e.g. www.Kazaa.com; www.neurogrid.net; http://edutella.jxta.org/; www.groove.net and Hyperclip, Sato et al. 2002].

One aspect of this that interests us is how this vision is beginning to spark new ideas for sharing information. This includes the idea that, with an owner's permission, you might be able to look into and use files from your peer's PC. For example: "most of the files in today's companies are on PC's, not servers, and peer-to-peer can let you see all these storage assets as one big distributed file space. A workgroup member might even be able to find the sketch of an idea you've just begun on your PDA" [Breidenbach 2001, para 26]

The idea of allowing others access to your "workspace", to work in progress, and to unique documents labelled and organised in ways that support personal use, is fundamentally different to the successful Napster–like models that have been used to share completed, static, often commercially created, predictable content, that has been specifically moved into a folder for sharing. As Bricklin [2000] argues, the reason Napster works is not merely that it uses p2p computing but that "the information being downloaded is never changed. The files shared with Napster are not news feeds – they are more likely the works of dead musicians".

There does in fact seem little justification for assuming that other types of personally owned content will be successfully shared through p2p computing purely because of the success of Napster. Rather than leap to that conclusion, however, we wished to explore these issues more systematically.

1.1 Approach and Focus

Although several groups are developing new tools to facilitate the sharing of other kinds of content (such as Edutella, Groove and Hyperclip), the focus of our

research was led by the questions being raised by groups from our own laboratory. For example, we have been working on new concepts intended to allow people to more easily share the benefits from the information they gather and organise from the Web [Banks et al. 2002]. Such tools would allow people to seek out peers with similar interests or expertise and to learn from the information they gather and use.

This then raises questions such as: What kind of information do people gather from the Web? How is it used? How is it kept? How is it modified? What aspects of it might be usefully shared with others? By examining these issues, we hoped to uncover both opportunities and obstacles in developing such systems, and to shed light more broadly on the issues that people developing p2p information sharing systems must consider. Our approach was to begin by looking at how knowledge workers do this. Knowledge workers, by definition, are people who spend a great deal of their time gathering, analysing, modifying and creating content. We also know that the Web is a key resource from which information is gathered by these workers to be kept and integrated with the personally-owned content on their PCs [Sellen et al. 2002].

1.2 Existing Research

The literature does provide us with an overview of information gathering and sharing tasks as they are carried out by knowledge workers, although not generally with an eye toward the design of information sharing tools.

Web-based information gathering is defined quite specifically as using the Web to purposefully find and collate information around a specific topic or theme. It is an interesting activity in the context of knowledge workers because earlier work [Sellen et al. 2002] has shown that this is the main and most important kind of Web activity that they carry out. Such activities very often involve sets of questions, ill-defined questions, or questions that are formulated in the course of carrying out a task. Information gathering is very different from some of the other kinds of Web activities knowledge workers do (such as fact finding) being generally more time-consuming and complex.

Some of these findings have been supported more generally, [Bates 1989; Hearst 1999; Pirolli and Card 1995; Turner 1997; Markus 2001; Paepcke 1996], the literature showing that information gathering is an iterative process, involving changing goals and the use of multiple sources, to gather together new ad-hoc collections of information. Those studies that have concentrated on the Web have indicated the advantage of domain knowledge, Web searching skills and the reuse of previously discovered sources upon the efficiency and effectiveness of gathering [Hoelscher and Strube 1999; Wexelblat and Maes 1999]. Resulting information can end up fragmented, residing in different formats and places, [Jones et al. 2001; Kamiya et al. 1996], not only on PC's but on paper, [Harper 1998], and in knowledge workers' heads, [Kidd 1994]. Kidd also suggests that the seemingly chaotic organisation of information during this process, which shows large individual differences [Berlin et al. 1993], is personally meaningful and allows the owner to use information, to be informed by it and to gain "knowledge" which may then be incorporated into new documents.

The methods by which individuals share the resulting knowledge, skills and gathered information from such tasks with others has also been studied, [Paepcke 1996; Berlin et al. 1993; Jones et al. 2001; Markus 2001; Wexelblat and Maes 1999; Bannon and Bødker 1997]. This literature indicates that information sharing is different and easier between close work colleagues or those who have shared knowledge and purpose than between loosely coupled colleagues, novices and experts or those who wish to reuse information for other purposes. Sharing between individuals is often observed within organisations or disciplines and it has been argued that information shared in this way preserves shared context and interpretations in a way that information shared through a central knowledge base, accessible to a wider audience, does not [Bonifacio 2001; Iamnitchi et al. 2002].

There have been a number of applications developed to support the sharing of information gathering processes [Wexelblat and Maes 1999] as well as sources, content and products [Kamiya et al. 1996; Takeda et al. 2000]. They may also support the discovery of individuals with similar interests or purposes and actively alert users to relevant information shared by others [Takeda et al. 2000]. Amongst the criticisms of some of these applications is that they may force the sharer to carry out extra work such as organising information into a different structure [Bonifacio et al. 2002]. This contrasts with the claims that p2p systems actually offer the opportunity to reduce the work done by sharers by allowing them to continue to gather, organise and use information using their own familiar tools and workspaces yet allow others to access this information with little or no extra effort being required [Kanawati and Malek 2000].

However Markus [2001] and Bannon and Bødker [1997] suggest that unless effort is taken in documenting information in a way that is reusable by different types of users for different purposes, others may have great difficulty in reusing that information. The implication is that enabling people to make information from personal workspaces easily accessible to others may or may not be of value.

It is clear from the literature therefore that these tasks can generate a multitude of documents throughout the process we define as information gathering. Yet we know little about what is shared, how it is shared and to whom it is of value. The literature suggests that work needs to be done in order to make information shareable and that the degree of work may depend upon the intended audience. Yet, we know little about what this work is or whether this work has already been carried out on personally owned information. We therefore aimed to investigate the potential for sharing personally owned information, in particular knowledge-based products, by studying what knowledge workers keep or create on their PC's as part of their work, what they currently do and don't share with regard to their personal stores of information, by looking at the way in which they share, and by looking at whom they share with.

2 Method

An exploratory approach was taken, capturing a rich amount of data using retrospective "walkthrough" interviews. Although some basic summary statistics

were carried out, the data were primarily analysed qualitatively using thematic analysis [see Aronson 1994].

2.1 Participants

Participants were recruited through email advertisements distributed via local mailing lists. These asked for knowledge workers, defined as people whose paid work involves significant time gathering, finding, analysing, creating, producing or archiving information, where "information" is anything from documents to drawings to multi-media files. From these respondents, 16 different knowledge workers were selected, across a diverse range of knowledge work, who were regular users of the Web for their work tasks. Regular Web use was defined as use of the Web at least 4 times in a typical working day.

Overall, participants had an average of 6 years of Web experience (ranging from 2 to 11 years), 4.5 years of experience of Web information gathering (ranging from 9 months to 10 years) and 6.5 years of experience in their current professional domain (ranging from 1 to 17 years). The resulting pool of people is summarised in Table 1.

No.	Job Title	Age Range	Yrs on Web
1	Customer Support (IT)	35-44	8
2	Information Resource Manager (Charity)	35-44	6
3	Education Officer (Charity)	25-34	10
4	Network Support Analyst	25-34	7
5	Territory Manager (Sales)	25-34	6
6	Development Manager (IT)	25-34	5
7	Games Producer	35-44	5
8	Graphic Artist	25-34	2
9	Architect	25-34	5
10	Lecturer and Union Representative	45-54	8
11	Government Policy Advisor	35-44	4
12	Building Historian and University Lecturer	55-64	4.5
13	Research Scientist	25-34	11
14	Government Planning Manager	25-34	2
15	Information Research Analyst	25-34	6
16	Researcher	35-44	6

Table 1. Summary description of participants.

2.2 Procedure

Each participant took part in a videotaped interview at their workplace, in front of their PC. Having been given the definition of information gathering [Sellen et al. 2002], they were asked to identify five or six information gathering tasks using the Web from the past couple of weeks (using their history list if needed).

Participants were then asked to verbally "walk-through" at least two of their tasks, most in fact covering more than this. Each participant started off by explaining the task they had carried out from how, when and why it was initiated

up until how it was completed (or up to the current point). They were also asked to open up browsers, bookmarks, email, paper folders and so on to show how they had extracted, created or moved information in each task. Not only did this support participants in recalling their tasks, it also provided additional data such as paper print-outs. Asking to be shown the artefacts being described also provided a simple, if crude, method by which the reliability of the retrospective accounts could to some extent be checked. Participants were also prompted with questions during the interview to elicit discussion about what they did and why, such as: Was this a typical task?; Where did the information come from and how was it found?; What was extracted, how was it used and why?; Was anything saved, recorded or created?; Has this been shared or could this be shared?; Where, how and with whom was it shared?; Would it be useful to reuse anything?.

2.3 Data and analysis

Data, in the form of videotaped interviews, were transcribed with the addition of notes concerning the artefacts shown to the interviewer and captured on video or paper (e.g. bookmarks, Web pages, printed documents). This material was then analysed task by task using thematic analysis. This involved categorising the data using two broad themes or frameworks, driven by the literature, the data and the research aims. The first, "the lifecycle of information gathering" (consisting of starting points, browsing and reading, extracting, storing and archiving and reuse) reflects the background literature and is largely based upon Turner [1997]. The second, "information sharing" (consisting of motivations and barriers to sharing, recipients and methods of sharing, and the work to make information shareable) was largely derived from the data. Together, the analysis within these two frameworks provides both an overview of information gathering tasks as well as the detail regarding the focus of the research: information sharing. Consistent with the analysis, the findings are presented within these two frameworks, presenting comments, behaviours, artefacts and so on that are related to particular stages of the lifecycle and to the components of information sharing.

3 Findings

Overall, 120 tasks were collected from the 16 participants (an average of 7.5 tasks per person, ranging from 3 to 14). Time spent doing these tasks ranged greatly from 15 minutes to 6 hours a day depending upon the stage of a project.

The majority of tasks (94) were examples of information being gathered for a specific current task such as gathering materials for a children's workshop, preparing a talk for a conference or getting ideas for a new computer game. However, some of the tasks (26) involved gathering information to satisfy a more ongoing interest such as regularly searching for organisations with similar interests, keeping up to date with what competitors were doing, or gathering illustrations or articles on a particular subject.

Unsurprisingly, while the Web was our central focus, in reality it was often one of many resources called upon in these tasks. In many cases, information was

gathered from other people and other document sources such as books, magazines and journals. Having said that these knowledge workers tended to rely heavily on the Web citing quick and easy access to a vast repository of information as now essential to their current work practices.

3.1 The Lifecycle of Information Gathering

Before looking more closely at the issues of sharing, we need first to look in more detail at how these information gathering tasks were carried out, or what we might call the "lifecycle" of this kind of process. Some interesting trends emerged when we looked at where participants started their search and where the products of these tasks ended up.

3.1.1 Starting Points

For most of these tasks (76 out of 120), participants started off with known Web sources (e.g. an organisation's website, an online database or a specific newsgroup) as opposed to Web search engines (used exclusively in 29 tasks), although sometimes they used a combination of the two (occurring in 15 tasks). A known source is a website that the participant may or may not have visited before but knows is there. They may know about sources through previous Web searching, through word of mouth recommendation or by anticipating that familiar real world sources such as people, publications or organisations will have an online presence. Comments suggested that through experience of use participants learnt about the information in a source, the domain and topics covered, the quality and accuracy of the information and the ease with which this information could be accessed. We could tell that at least half of the known sources had been visited before because they were accessed via bookmarks or self-authored Web pages.

With regard to search engines, participants used these either when they could not think of a useful known source or when they tried and failed to find the information they needed. They also tended to go straight to search engines when the topic of information was unfamiliar. What the data show then, is that these knowledge workers more often than not stayed within familiar domains and used familiar resources to begin their information gathering tasks. Knowledge gained about particular sources was reused in order to find and select a starting point for a task.

3.1.2.Browsing and Reading

Once a Web task had begun, participants looked through many different kinds of information, seeking information not only relevant to the topic at hand, but also anything they found new, interesting, comprehensive, accurate, up to date and well presented. This process almost always involved multiple sites, and could take place over hours, days or even weeks.

One interesting aspect of this was that the learning was often in the gathering. Many participants talked about picking up knowledge throughout the whole process of information gathering such as gaining background information, getting

to know important keywords, and learning specific pieces of information as they went from site to site. It was common to hear the study participants talk of starting their searches wide to understand the bigger picture of a topic before focussing in on detail:

"I start off fairly wide and then hone it down to particular events so then if I find something useful, started off at the 1750's, got 1700's timeline, particularly got interested in slightly later,…….and then I do a search on (name of historical event) and hone it down so you've got information, quite a lot of information on particular, literally a particular day if possible" (Games Producer)

Not only were they picking up domain knowledge through this process, they were also developing their search strategies and skills. Comments suggested that such skills were perceived as essential in being able to carry out knowledge workers' work effectively.

3.1.3. Extracting.

In addition to the implicit process of information extraction that went on in almost all of the tasks, participants also explicitly extracted pieces of information from the Web by copying and pasting into documents, saving whole documents as files, printing, bookmarking, archiving in email, making written notes or saving in personalised Web folders.

For example, the Customer Support person typically sought advice both from colleagues (via email, phone and face-to-face conversations) as well as searching the Web in order to find solutions to customer problems. Good sources of Web information would be kept as a bookmark possibly later being incorporated as a link on his personally authored intranet page. In addition, gathered information from various sources would sometimes be copied and pasted into a Word document. This document might be emailed to a person who would place it either on the intranet or on the Web depending on his instructions. Associated email messages were kept including the attached documents. In addition, he often saved many downloaded files or patches from the Web on his hard drive. Those that he was able to distribute would later be moved to the server for his colleagues to access.

As this illustrates, in addition to the information kept in the heads of the knowledge workers, any particular information gathering task could have associated with it several different informational "by-products" in different formats, residing in different places. Any of these by-products could be reformatted or otherwise modified or transformed more than once. Some of these were transient or temporary, and others were useful in and of themselves. Some of these by-products were shared and others were not. Understanding how they are related and where they have come from may be quite complex.

3.1.4. Storing/ Archiving

Figure 1 provides a snapshot of where these by-products ended up in participants' own information spaces. For example, Urls most commonly ended up in bookmarks, 'Other' sorts of information (such as text or images) most commonly

ended up in the documents kept in personal folders. In general it was more common for information to end up in "personal spaces" that were only accessible to the participant (i.e. personal, email and bookmark folders on the hard drive or personal folders on the network), than for information to end up in "shared spaces" that were also accessible to others (i.e. intranet pages, shared network folders or public Web pages). It was also evident from both comments and observation that these personal and shared spaces differed in both content and organisation.

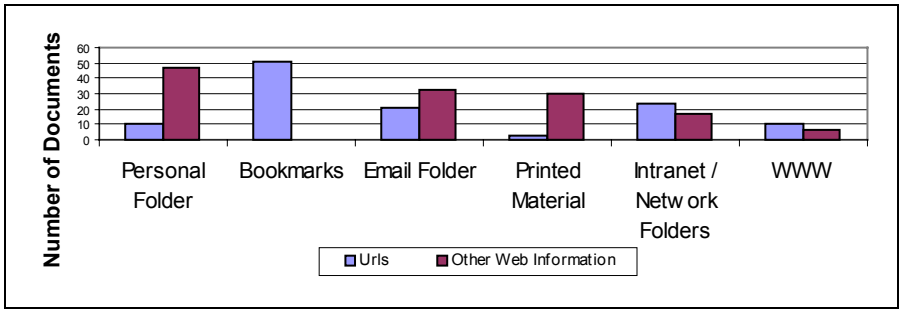

Figure 1. Graph showing the number of 'documents' containing either Urls or other extracted Web information ending up in various different storage places

With regard to content, information stored in personal spaces was described both as "personal" (i.e. non-work) and as work information that was not "useful" or "relevant" to anyone else. Information may be held temporarily, in draft form or be being kept as a record of a task that has been done. Personal spaces can also be used as a dumping ground for information that does not belong anywhere else or is not appropriate to put in a shared space. Personal non-work information does not tend to be hidden away, a concern should other people suddenly have access to a personal space.

With regard to organisation, although personal spaces may be described by participants as fairly organised by topic or project, it was pointed out by one individual that someone else trying to use this information would at least need to know "what I was doing and what I was supposed to be doing" (Games Producer). Indeed there was evidence of "cheating a lot" (Architect) when it came to personal organisation, in that information that strictly did not "fit into" a folder's category may be put there and similarly information that could be filed was not. By contrast, shared spaces demanded more consistent organisation so that others could find information easily. Multi-contributor shared network spaces meant things might be more "tricky" to find. In one case contribution to a shared database was controlled. This was not to do with the organisation but to do with controlling the quality and amount of information that was shared, illustrating other factors are also important.

What this suggests is that participants utilised different spaces in different ways. One consequence of this is that the information in personal and shared spaces differs in its content and organisation.

3.1.5. Re-use

A final issue which interested us was the potential reusability of the resulting collection of Web-derived information on gatherers' desks, PCs, and networks. Here the findings were quite striking. While participants often reused sites and sources as starting points, in only 3 of 120 cases was any content or were any documents from past projects reused. In addition, when asked, participants said they expected to reuse information in future projects in only 12 of 120 cases.

It was quite clear then that these knowledge workers were creating bespoke products on a project-by-project basis. The way information was gathered, extracted and modified was done for the specific purpose to hand, and that purpose changed with each new project. As the Education Officer put it:

"[The Web] is a good base of resources, …[you] will want to take pieces of it …. It's not even a jigsaw, like cooking almost, you take all these relevant bits and you mix them together to make your own recipe".

By contrast, what these knowledge workers were reusing were the sources of information (and the knowledge of how to find them) together with the skills of gathering information, something they learned from long experience. For example, the Games Producer described methods of searching and extracting information that he used again and again over years of doing research. As he put it:

"Its only when I see somebody who hasn't had [my] background try to research something that I find out that actually I'm quite good at this".

3.2. Information Sharing

Turning from the lifecycle of such tasks, we now look closer at the sharing of information in such tasks.

3.2.1. Motivators and Barriers to Sharing

When we asked participants about their initial intentions, in 71 of the 120 tasks (60%) they said they were expecting to share some part of the output of their tasks. About two-thirds of these cases were driven by obligations (such as a request for information or an expectation on both sides that information will be shared, often laid down by routines or work practice). In the remaining third, participants intended to share with recipients who were not necessarily expecting anything from them. Reasons for this self-initiated sharing were often to do with promoting oneself or the organisation, placing work obligations onto someone else or informing the recipient of something they ought to know. In addition to these 71 tasks where participants expected to share, there were 9 further cases in which the intention to share developed during the task (afterthought sharing).

Interestingly, there were many factors involved in why some information was not shared. In some cases, participants were restricted by copyright or company confidentiality restrictions. In other cases, participants wished to keep personally relevant information confidential (e.g. the Territory Manager was concerned that some of his bookmarks revealed his interests and where he banked or shopped).

Some participants were unsure as to whom might find it useful (e.g. the Researcher copied information into an email to share with work colleagues and then deleted it, not being sure whether it would be useful or interesting to colleagues). Additionally not knowing how to share, or the effort involved in sharing, could influence whether something was shared (e.g. the Information Resource Manager wanted to share bookmarks but could not remember how to do this. And similarly, the Research Scientist said he would have liked to share more information on his Web page but having to pass this information through a colleague to make the alterations, had not got around to doing so).

3.2.2. Recipients and Methods of Sharing

Sharing was most often with individuals or small groups, and this was done mainly through methods that delivered information direct to the recipient(s) (i.e. via email, fax, memo or face-to-face). On fewer occasions, sharing was with the larger organisation or the public, this being done mainly via central repositories (the Web, Intranet or central database or store). This is illustrated in Figure 2.

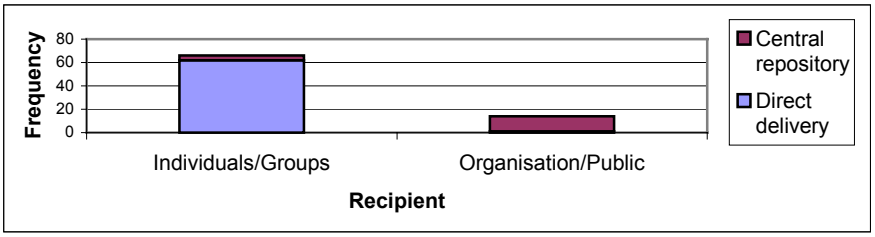

Figure 2. Graph illustrating how often information was shared with individuals/ groups versus organisations/general public.

3.2.3. The Work to Make Information Shareable

One of the issues we were most interested in was what, if anything, tended to be done to the extracted Web information in the case of information intended to be shared versus that not intended for sharing.

We found that information intended for sharing was *always* modified after it was extracted, whereas this was very unlikely to occur in information that was not shared (Figure 3). Specifically, we found that shared information could be modified three ways, none of which was mutually exclusive. It could be: (i) rewritten, (ii) written around, or (iii) enriched at the point of delivery through the attachment of extra information. In this last case, we mean that participants talked of adding context to, or explaining the information they were delivering through conversations, e-mail, faxes or memos at the point of handing it over.

The important point to note here is that shared information was modified in ways that non-shared information was not. This is not to say that no work was associated with personal information, but rather that this work was largely mental work (e.g. reading and comprehending), or work at the point of extracting the information

which was limited to filtering, categorising and organising the information as opposed to modifying its content in any way.

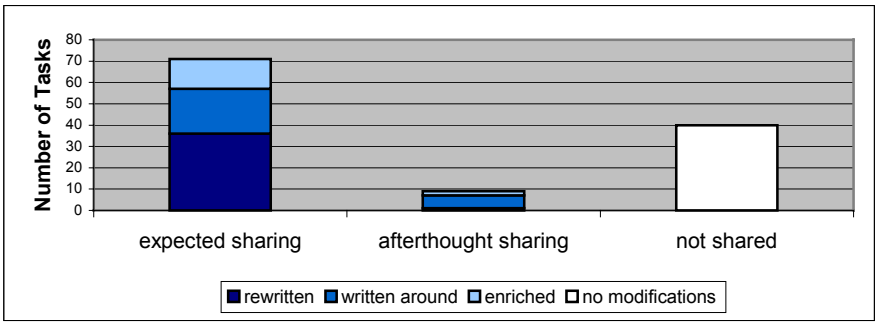

Figure 3. Number of tasks in which people shared information as expected or as an afterthought or in which information was not shared. Also shown is whether the information was rewritten, written around, enriched or unmodified.

So what exactly was being accomplished through modification? Looking more closely, we can see that there were many ways in which the information was being re-designed to make it easy to understand and to use by its recipients. This is illustrated by a quotation from the Education Officer:

"you know teachers have not got time for anything really .. and they keep getting told to do new things and so if you're making it easy, they're going to use it. So I mean especially for education, it's a really useful tool, the internet, ... information is what I am in the business of, information finding, it is an absolute nightmare ...there must be ways of making it more useful for them (teachers)".

While the idea of recipient design is certainly not new, [going back to Harvey Sacks in the 1960s; see Sacks 1992], it is interesting to look at what this means in terms of sharing Web information. There were many ways in which this took place:

Checking and filtering: During the gathering process itself, participants checked and filtered information to ensure its usefulness. Information was checked for accuracy against other sources (e.g. experts, other Web pages, own knowledge or colleagues), and judgements were made as to whether the information was of good quality and up to date. Participants also talked of only sharing information from sites that were trusted or familiar.

Translating, modifying and organising: After extracting information, participants changed or re-organised information to ensure that it was easy to find, use and understand. This was done by:

- Changing the information format, size or language. File formats were changed to those that the recipient was most likely to be able to access or use (e.g. from CAD to bitmap; from digital to print). File sizes were cut (by zipping, removing content or reducing resolution of images). There were even examples of translating content into a recipient's native language.
- Simplifying. Information was sometimes simplified to suit its readership and a recipient's concerns. For example, language was simplified for use

by children or information was cut down to its bare essentials, to reduce time and effort in scanning information, and to make it relevant to a recipient's request;

- Guiding by highlighting, organising, signposting and explaining. Several types of cues were added to information to guide the reader. This ranged from highlighting important information, to organising and structuring information in clusters, to inserting headings and labels. Sometimes, this included adding more explicit "signposts" (such as step by step instructions on how to navigate the information or by adding descriptions of which link to choose for what). Sometimes it included overviews or summary explanations telling the recipient what the information was for and why they were sharing it.

Enriched delivery: We also saw that added messages, explanation or discussion often took place at the point of delivery, even if this was not done physically but in the digital realm (e.g. through email). These findings are consistent with other studies that show that information, such as documents, are often discussed with the recipient at the point of delivery [Harper, 1998] in order to put them in context.

Maintaining and updating information: Finally, even after information was made available to others, sharing information on a persistent basis (either through html pages or regular email bulletins) also placed a burden of maintenance upon the participant. This meant there was a need to regularly add new information, update old information and check links on Web or intranet pages.

Therefore our participants expected to and did share information in a substantial proportion of their information gathering tasks despite a variety of factors that could act as barriers to sharing. Information was usually shared with individuals or groups using methods that allowed information to be delivered directly to the recipient(s). Although some form of work was carried out on the information gathered in the tasks, the work carried out on shared information was fundamentally different in nature and extent when compared to that carried out on non-shared information.

4. Implications and Conclusions

To conclude, there are several key points to take from these findings, each of which has implications for p2p information sharing:

- *The knowledge workers in this study more often than not dealt with familiar topics and used familiar resources to begin their information gathering tasks.* By implication, we can assume that since knowing the source of information is so important in the Web gathering process, then likewise, p2p knowledge sharing systems will need to make explicit some important aspects of the source of p2p information. For example, it may be important to know aspects of the person from whence the information came (e.g. what their expertise is) as well details about how and from where they gathered their information.
- *Knowledge workers learn from the process of searching and gathering.* By implication, if users are given access to the products of someone else's search efforts (essentially bypassing their own search and gathering processes) it is

conceivable that they might also bypass the opportunity to do some of their own implicit learning. Access to the products of someone else's work might therefore ultimately be shown to be a less effective way of gaining knowledge than doing one's own information gathering work.

- *Any given information gathering task may have associated with it many different kinds of informational by-products. Some are transient or intermediate products and some are more properly "end " products. Some are Web-based and some are not. Some exist in the digital world and some are physical. Thus, by-products of these kinds of tasks are part of an 'information ecology' where the relationship between artefacts over time and space may be important to understand.* The implication here is that looking at any one document or piece of information may not be useful without looking at the bigger picture, or without understanding how it has come to be. As pointed out by Kidd [1994], this contextual knowledge may be in the head of the person who has created the information, but may not necessarily be obvious to an outsider looking "in". This is related to the next point.
- *Personal and shared spaces are used differently, the content and organisation of the information being found in personal information spaces being fundamentally different to that being kept on shared spaces such as the Web. Personal spaces are akin to "workbenches" whereas shared spaces are more like "archives".* The implication is that p2p systems that tap into personal spaces are, as predicted, likely to unlock information that is fundamentally different to that found in archives, central repositories or the Web. By their very nature they may not be browseable or searchable in the same way as archives are. Therefore traditional p2p mechanisms for file sharing may not be appropriate for sharing information in workplaces. This is aside from questions raised in the following points as to whether this content is actually of any value to other people.
- *Information that has been gathered with a specific task in mind is rarely repurposed or reused for new tasks.* This then begs the question: If information gathered by an individual is not perceived to be usable in the future by its owner, to what extent can any of these products be reused or repurposed by other people? By implication they will have to have similar purposes or tasks to hand. This suggests that p2p tools need a way of allowing users to effectively specify and match these across users.
- *Knowledge workers often shared information as part of their information gathering tasks and this was mostly to individuals and groups using methods that delivered information directly to the recipients.* This implies that at least some products from these tasks are usable and shareable with peers [which is consistent with the findings that information tends to be shared within small communities of interest; Iamnitchi et al. 2002]. However, p2p systems may benefit from looking at ways in which information can be directly delivered to recipients in response to an explicit request or through identifying recipients who have an interest or need for the information. This would be more consistent with the way that knowledge workers currently share information as opposed to models where information is placed in an area such as the Web for people to find and gather for themselves.

- *The work carried out on non-shared information is not the same in nature or extent as the work carried out to make information shareable, where information is prepared with specific recipients or audiences in mind.* Given some information will have been prepared for sharing and others will not have, it will not be obvious how to distinguish between the two let alone identify within a workspace those documents which match the needs of any particular recipient. Designers of p2p knowledge sharing tools might usefully learn by looking at the kind of work that is done to make information shareable by others. Some of this work might be done automatically, but some may not. It is likely, given the nature of these modifications, that the experience gained in the process of gathering, using or creating a knowledge-based product in turn provides much of the knowledge required to effectively prepare this information for sharing. In any case, merely enabling access to another person's information will not be enough to leverage the knowledge between them.

By studying knowledge workers' Web information gathering tasks we have highlighted both constraints and opportunities for the design of p2p information sharing systems. This work can be treated not only as a set of cautionary notes about some of the underlying assumptions of such systems, but also as pointing towards ways in which such tools might be developed in new ways. While we have focused on web-based information gathering, we believe these results have more general application for the sharing of information across individuals, whether it is derived from the Web or not. Ultimately, the generalisability of these data, as well as specific design solutions will depend on a more extensive programme of user research than is reported here.

Acknowledgements

We thank all the knowledge workers who took part in this study both for their openness and their insights. Thanks also to Richard Harper and Martin Merry for comments on an earlier draft.

References

Aronson, J. [1994], A Pragmatic View of Thematic Analysis. *The Qualitative Report*. 2 (1). At: http://www.nova.edu/ssss/QR/aindex.html [April, 2003].

Bates, MJ. [1989], The design of Browsing and Berrypicking Techniques for the on-line search interface. *Online Review, 13(5),* 407-431.

Banks, D., Cayzer, S., Dickinson, I. & Reynolds, D. [2002], The ePerson Snippet Manager: a semantic web application. *HP Labs Tech Rep HPL_2002_328.*

Bannon, L & Bødker S. [1997], Constructing Common Information Spaces. *ECSCW '97conference proceedings.* 81-94.

Berlin, L.M., Jeffries, R., O'Day, V., Paepke, A., & Wharton, C. [1993], Where Did You Put It? Issues in the Design and Use of a Group memory. *HP Labs Tech.Rep HPL-93-11.*

Bonifacio, M., Bouquet, P. & Traverso, P. [2002], Enabling Distributed Knowledge Management. *Informatik.Informatique. 1,* 23-29.
Breidenbach, S. [2001], Peer-to-peer Potential. *Network World Fusion*: http://www.nwfusion.com/research/2001/0730feat.html , [April, 2003].
Bricklin, D. [2000], Thoughts on Peer-to-peer: http://www.bricklin.com/p2p.htm [April, 2003].
Harper, R. [1998], *Inside the IMF*. London: Academic Press.
Hearst, M. [1999], User interfaces and visualization. In R. Baeza-Yates & B. Ribeiro-Neto (eds) *Modern Information Retrieval.* ACM: NY. Ch 10: 257- 323
Hoelscher, C. & Strube, G. [1999], Searching on the Web: Two types of expertise. (Poster extract) *SIGIR '9:* 305-306.
Iamnitchi, A., Ripeanu, M. & Foster, I. [2002], Locating Data in Small-World: Peer-to-Peer Scientific Collaborations. *1st International Workshop on Peer-to-Peer Systems, MIT, March 2002*.
Jones, W., Bruce, H. & Dumais, S. [2001], Keeping Found Things Found on the Web. *CIKM'01, Nov 5-11.* 119-126.
Kamiya, K., Röscheisen, M. & Winograd, T. [1996], Grassroots: A System Providing a Uniform Framework for Communicating, Structuring, Sharing Information, and Organizing People. *Computer Networks and ISDN Systems.* 28: 1157-1174.
Kanawati, R. & Malek, M. [2000], Informing the design of shared bookmark systems: http://citeseer.nj.nec.com/414039.html , [April, 2003].
Kidd, A. [1994], The Marks are on the Knowledge Worker. *CHI '94:* 186-191.
Kubiatowica, T. [2003], Extracting Guarentees from Chaos. *Communications of the ACM.* 46 (2): 33-38.
Markus, L. [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.
Paepcke, A. [1996], Information needs in Technical Work Settings and their Implications for the Design of Computer Tools. *CSCW*: 63-92.
Pirolli, P. & Card, S. [1995], Information Foraging in Information Access Environments. *CHI '95.* ACM: New York, 51-58.
Sacks, H. [1992], *Lectures on Conversation, Vols 1 and 2.* Edited by Gail Jefferson. Oxford: Blackwell.
Sato, H., Abe, Y. & Kanai, A. [2002], Hyperclip: A Tool for Gathering and Sharing Meta-data on Users' Activities by Using Peer-to-peer Technology: http://www.cs.rutgers.edu/~shklar/www11/final_submissions/paper12.pdf ,[April 2003]
Sellen, AJ., Murphy, R. & Shaw, K.L. [2002], How Knowledge Workers use the Web. *CHI 2002, April 20-25.* 4(1): 227-234.
Takeda, H., Matsuzuka, T. & Taniguchi, Y. [2000], Discovery of Shared Topics Networks among People – A simple approach to find community knowledge from WWW bookmarks. *PRICAI2000:* 668-678.
Turner, K. [1997], Information Seeking, Retrieving, Reading and Storing behaviour of Library-Users: http://citeseer.nj.nec.com/264174.html , [April, 2003].
Wexelblat, A. & Maes, P. [1999], Footprints: History-rich tools for information foraging. *Proceedings of CHI '99.* New York: ACM Press: 75-84.