Gathering and sharing Web-based information: Implications for “ePerson” concepts

Jennifer Hyams, Abigail Sellen
Mobile and Media Systems Laboratory
HP Laboratories Bristol

HPL-2003-19
February 6th , 2003*

E-mail: jenhya@hplb.hpl.hp.com, abigail.sellen@hp.com

ePerson, peer-to-peer, web use, knowledge workers, collaboration, information re-use, knowledge management, sharing, information gathering

“EPerson” is an example from HP Labs of the kind of new ideas for knowledge management and sharing that are being inspired by Peer-to-Peer and Semantic Web technologies. To help inform and ground the efforts of such developments we report on a study looking at how people gather and share information using today’s digital tools. We look at 16 knowledge workers’ use of the Web, examining the lifecycle of this information and the way in which it is shared. We find that Peer-to-Peer knowledge sharing is different to say Peer-to-Peer music sharing, user’s PC’s being more like “workbenches” than archives. We also find that people learn through the process of gathering, creating and organising their information. Such knowledge enables individuals to reuse and share their information, personally owned information always being modified or enriched before sharing. Such insights provide important implications for developing Peer-to-Peer file sharing technologies which we discuss.

Gathering and Sharing Web-based Information: Implications for “ePerson” Concepts

Jennifer Hyams and Abigail Sellen

Hewlett-Packard Labs
Filton Rd., Bristol, BS34 8QZ UK
jenhya@hplb.hpl.hp.com, abigail.sellen@hp.com

Abstract

Recent interest in Peer-to-Peer computing and the Semantic Web (e.g. Berners-Lee, Hendler and Lassila, 2001) has inspired new ideas for ways in which people might gather, manage and share knowledge through networked tools and infrastructures. In particular, the “ePerson” group at HP Labs Bristol is researching and developing ways in which new Web-based structures and applications might provide useful intermediaries or electronic representations for people in networked environments. An electronic representation of a person might, for example, help a user control and protect their identity on a network, help seek out information more effectively through a network, and help leverage knowledge from others more efficiently. In order to help inform and ground the efforts of such developments we report on a study designed to look closely at how people gather and share information using today’s digital tools. This is done by looking at 16 knowledge workers’ use of the Web, examining the lifecycle of this information and the way in which it is shared. We find that Peer-to-Peer knowledge sharing is different to say Peer-to-Peer music sharing, user’s PC’s being more like “workbenches” than archives. We also find that people learn important skills and knowledge through the process of gathering, creating and organising their information. This knowledge enables individuals to reuse and share information. In particular we found that personally owned information was always modified or enriched before sharing. Such insights into how information is reused and made sharable provides important implications for developing Peer-to-Peer file sharing technologies which we discuss.

Keywords

ePerson, Peer-to-Peer, Web use, knowledge workers, collaboration, information re-use, knowledge management, sharing, information gathering

Introduction

This report describes a study carried out in collaboration with the “ePerson” group in HP Labs Bristol. EPerson is a concept that provides an “active representative of an individual on the Net”. It is a Web service that holds information for individual users, placing them in control of the information that they possess, whether this is information about their owners (such as personal profiles) or information that they have gathered or created. It is envisioned not only as a personal information management service but also as an intermediary, acting on a person's behalf to exchange information with others on the Net. In this role, a user's ePerson represents the user's wishes and interests in terms of who is allowed to gain access to their information and what information is shared. In negotiating such exchanges, the ePerson could similarly gain access to information from others in line with its owner's needs and interests. (For more information on ePerson see Banks, Cayzer, Dickinson and Reynolds, 2002).

There are many potential benefits for such an intermediary including: controlling privacy and anonymity, proving identity, personalising and speeding up login procedures and interactions with websites, automating or supporting tasks such as planning trips or meetings, collaborating with other ePersons to obtain information relevant to their owner's needs or interests, identifying other individuals with shared interests, maintaining personal history information and so on.

EPerson therefore covers a range of different concepts and possible application areas. However, the focus for this study was on personal information management and the sharing of information between peers (as represented by each person's ePerson). Consider the following example:

> Imagine "Sam". Sam wants to write a report on the trends in cult TV programmes and is looking for information to help in this task. Sam goes to Google, types in some keywords and gets several thousands of search results from the Web to start scanning and browsing through. Sam opens up some of the pages that look relevant, scans a page and copies and pastes a section of the text and an image into Word and carries on browsing. Sam often returns to Google to refine the search. Several hours later, Sam has a collection of bookmarks, printed pages and saved documents on the PC that later will be read in more detail.
>
> Imagine instead that Sam has a different way of finding this information, one that leverages the work already done by Sam's peers. Sam opens the "community browser" and types in the query "Cult TV". EPerson searches the information other members of the community have about Cult TV and filters the results according to Sam's preferences. It then presents a list of links to collections of information on Cult TV. Sam chooses a link, seeing that this collection is large, very recent and belongs to "Robin" an academic with an interest in this area. A collection of documents already gathered and organised on the topic of interest have their titles displayed on the screen. Sam can see a range of information in this collection: documents gathered from the Web, PDF files, images, emails as well as documents written by Robin. Sam browses the titles in this collection and can see how Robin has organised the information by types of programmes such as "Sci-fi" and "Children's" but has placed his own reports in another grouping. Sam "flags" the Sci-fi collection and all of the reports Robin has written that

> have used information from the Sci-fi collection are highlighted. Sam also "flags" these reports and these together with the Sci-fi collection are downloaded to Sam's own Cult TV collection.

The way in which people gather, use, store and share information is pertinent to developing tools to support such scenarios. We believe that studying the first scenario of how people currently carry out this process can inform the potential for opportunities for using new kinds of Peer-to-Peer architectures and services in the second scenario. It can do this by:

- providing a rich set of real examples and scenarios of use to ground the development of new tools and infrastructure;
- providing a reality check against which some of the assumptions underpinning new ideas can be calibrated;
- helping to highlight problems that people currently have with the tools they have to hand;
- pointing toward the potential value of the technologies being proposed, helping to guide their design and development.

With these goals in mind, this report describes an in-depth study of how users gather, use, store, archive and share information with some of the current technological tools available. The main focus in this case is the Web, although we look at other kinds of tools (both digital and more mundane tools such as paper) that are used in conjunction with Web-based information gathering and sharing. In this study we explore the “what and how” of the whole information lifecycle from how information is initially gathered from the Web through to where that information ends up and what form it takes. On the basis of this set of data, we then derive a set of implications for new file-sharing tools such as ePerson.

We start then with an overview of the literature before we go on to outline our approach and present the body of our findings.

Literature Review

In this section we take a look at the existing literature on information gathering and sharing, and the tools that have been developed to support people in these tasks. We also provide a summary of relevant Peer-to-Peer computing research to illustrate the kind of technology that may be the bedrock of future tools in this area. Following this review, we then define our approach, which includes explaining our rationale for focusing upon the Web as an information source and knowledge workers as our sample population. Here we also detail the definition of “information gathering” we used to address the specific aims of this study.

The Information Lifecycle

In our coverage of the information lifecycle we look at gathering, using, saving/archiving and sharing activities. We are not suggesting that these are necessarily distinct separate activities. For example, the literature suggests that users read, assimilate and analyse information during the search process (O'Day and Jeffries, 1993) as well as after it. This contradicts Card et al (1996) who argue that the slowness of the Web means that it is hard to integrate finding information with "using" it, the implication being that finding information is a clearly defined separate task to using information. Similarly, the archiving of information may be viewed as a separate task for the user. Kidd (1994) makes this explicit by pointing out that filing is not a primary goal of knowledge workers and that in fact information cannot be properly filed until it has been understood and has informed the reader. In the lifecycle approach we wish to take, irrespective of the degree of integration between these tasks, these tasks are all part of the information lifecycle process and it is within this approach that we wish to gain a greater understanding of what happens to gathered information as opposed to studying these tasks separately. We cover each of these tasks in turn below.

Information Gathering

User studies of information gathering have led to new ways of thinking about how people seek out information. The traditional view was one in which a user had a stable goal throughout the search process, searching for end documents that met this goal. Many search interfaces were designed on the basis of this user model (Hearst, 1999). However, although these accounts incorporated the fact that users reformulated their search terms and carried out iterative searches (something shown by O'Day and Jeffries, 1993), they did not account for the findings that users were seen to learn during the process, to gather together new ad hoc collections of information throughout the search from multiple sources, and to shift both goals and queries during the search process (e.g. O'Day and Jeffries, 1993; Kuhlthau, 1993; Hearst, 1999; Paepke, 1996). Models that reflected these observed behaviours such as Bates' (1989) "Berrypicking" model and Ellis, Cox and Hall's (1993) "Information Seeking" model have therefore been receiving attention in the digital information search arena, both consequently being applied to interface search behaviour and being used to make design recommendations for searching interfaces for the Web and digital libraries (Turner, 1997; Bates, 1989; Hearst, 1999).

Additionally there is also a growing body of literature that has focused specifically upon Web search behaviour and patterns. "Information Foraging theory" (Pirolli and Card, 1995; Card, Pirolli, Wege, Morrison, Reeder, Schraedley and Boshart, 2001) has been developed particularly around Web search behaviour. It is a theory of why users navigate to the pages that they do. Their theory suggests that users optimise their time by seeking out patches or clusters of information and avoid those of perceived low value. They bring in the concept of "information scent", that is, cues (text, graphics etc) on a page that are used to predict the usefulness of content at the other end of the links from that page. Again this theory has been applied to redesigning search interfaces specifically for the Web (e.g. Web Forager and WebBook, Card, Robertson and York, 1996).

Other studies have looked specifically at Web navigation and surfing patterns. Tauscher and Greenberg (1997ab) identified seven Web browsing patterns; (first time visits to a cluster; revisits to pages; page authoring; use of Web-based applications; hub and spoke; guided tour; depth-first search). Wilson (1997) identified four categories of information-seeking, this time differentiated by the method by which information is gathered ("passive attention", "passive search", "active search" and "ongoing search"). Other studies such as Hoelscher and Strube (1999) and McKenzie and Cockburn (2001) also warn us of the other factors that may play a part in the behaviour and navigational path that a user takes, such as the advantage of domain knowledge, Web searching skills and re-using previously discovered sources upon the efficiency and effectiveness of gathering.

To complicate this picture further, it is also evident in the literature that "information searching" itself can be broken down into various categories, and indeed it has been. For example, O'Day and Jeffries, (1993) distinguish among monitoring a well known topic, following a plan for information gathering and exploring a topic. Choudhury and Sampler (1997) distinguish between reactive and proactive information gathering. The implication here is that the type of search or goal that the user has will influence the searching behaviour the user exhibits. For example, in monitoring tasks, users tend to go straight to the details, already having the background context (O'Day and Jeffries, 1993). Marchionini (1995) in a review of the literature identified three types of browsing differentiated by the type of goal or information sought: "directed browsing" – systematic and focused on specific objects or targets; "semidirected browsing" – predictive or generally purposeful; and "undirected browsing" – no real goal and little focus. Thus, although the models described above do not distinguish different patterns for different types of gathering tasks, the literature suggests that we should identify the types of tasks we are talking about because of the likelihood that different types of tasks will have an influence upon the exact nature, content and sequence of behaviour. Thus in this study we specifically define what we mean by "information gathering".

Some tools aim to simplify and personalise the information gathering process for individuals. "Mindheap" (Takashiro and Takeda, 2000) attempts to do this by learning the kinds of information its user is interested in and then recommending phrases or other pages to be viewed. "Siteseer" (Rucker and Polanco, 1997) attempts to do something similar, but uses personal interests to search the bookmarks of others in order to recommend webpages.

Using Information

Using information may involve extracting, sensemaking, modifying, organising and creating new artefacts. For example, Turner (1997) illustrates the printing, reading, listening or viewing of documents in order to extract information and the making of annotations and notes which makes information that can be used later on. Similarly, O'Day and Jeffries (1993) illustrate the work that information intermediaries do in reading and annotating, finding trends, making comparisons, aggregating information, identifying a critical subset, assessing and interpreting. This often results in the creation of new artefacts, summaries or overviews which are influenced by the user's domain knowledge and purpose.

Therefore it is important to note that information is being gathered for a purpose and this purpose or goal is likely not only to have an influence upon how information is gathered but also upon what and how it is used. Kidd (1994) provides an interesting example suggesting that different types of tasks, traditionally carried out by different types of workers, may influence how information is handled or modified. For example, a communications worker may collect and classify information, enrich it and pass it on. A clerical worker may collect and apply information (e.g. policies) and the information remains unchanged and the worker is not informed by it. In contrast, a knowledge worker processes, extracts information and produces new concepts, ideas, and manipulations of the information.There are therefore opportunities for the development of tools that are better tuned to particular usages and purposes. For example, Knight and Munro's (2001) "intelligence amplifying tools" use visual representations to specifically aid users in analysing data. Consequently, we define the characteristics of our participants in terms of the type of work they do as the literature suggests that this is an important factor to note.

Saving/ Archiving

Information obviously ends up fragmented, residing in different formats and places (Jones, Bruce and Dumais, 2001; Kamiya, Röscheisen and Winograd, 1996; Peters, 2001). Where information is kept at any one time may depend upon whether the information is still being used and how it is being used. The organisation of people's workspaces or personal information spaces often reflects this. For example, Card et al (1996) show how a small amount of the information people keep is organised to be available at low cost, moderate amounts at moderate costs, and large amounts at large cost. Sellen and Harper's (2002) notion of hot, warm and cold documents is similar to this, hot documents being kept close at hand for work in progress, cold documents being stored away. But they also emphasise the large individual differences in information organisation and the meanings inherent in the way documents are organised, often despite first impressions of chaotic desk spaces and piles of paper (a notion that also appears to apply to Web-based bookmarks, see Abrams, 1997). (See also Kidd, 1994; Malone, 1983; and Berlin, Jeffries, O'Day, Paepke and Wharton, 1993). For example, Sellen and Harper's (2002) studies of the organisation of information in folders showed a diversity of documents and organisational strategies that reflect how the information was being used. Ordering and interleaving documents, the use of annotations, post-its and so on all provide cues to the user that help to show the state of play and history of a user's activities and provide reminders of work in progress. They suggest that once information is archived away, many of these cues are lost being removed from the activities they once related to.

Kidd (1994) similarly warns that a knowledge worker becomes informed through the reorganisation of information on desks, the making of notes and the contextual cues that are available. These provide different outputs for different knowledge workers from the same knowledge. She argues that much of this "informing" is carried around in the person's head and is not residing within the documents although these thoughts may be translated into reports. She suggests that, at least for knowledge workers, once used, information discharges its value, having been read and understood and new artefacts created. Filing thus becomes secondary and effortful, not strictly being necessary in order for the primary knowledge task to be done.

Other studies have focused upon those documents that have gone "cold" or, in other words, those documents that have been archived away. Whittaker and Hirschberg (2001) looked at personal paper archives and found that office workers kept large and valued paper archives, 22% of which included obsolete information, low-value information, and things that had never been read. Of the remaining information, about half were unique documents (usually written by or for the archiver) the rest being publicly available and unread. People retain information for perceived future value and, once filed, are likely to keep the document because of the effort involved in later finding and discarding information that has little value. Indeed, people may forget their filing categorisations and create duplicate files. In terms of the future use of these documents for themselves or by others, this study suggests that a proportion of the information kept has lost its value to the filer, a lot of information being kept "just in case". Some information will also be being kept for sentimental or personal reasons and will be of little interest to others. However, people are seen to keep a proportion of documents of high quality that influence their work. They also found that people rely on others to collect information on certain topics rather than duplicating this in a personal information archive, illustrating both that some stored information is perceived as reusable and as sharable, and that information collected by peers can be perceived as a useful source of information.

Archives need not necessarily be "personal". In some organisations, centralised, shared repositories are used with the aim of facilitating the sharing of knowledge. However, there may be several barriers to sharing information in this way, such as lack of personalisation. Users may have to organise information into a different structure to that of their own information space (Bonifacio, Bouquet and Traverso, 2002). Furthermore, agreeing upon a shared categorisation or organisational structure is a difficult task (Sellen and Harper, 2002) given that information can fit into multiple folders and be categorised in many ways (Abrams, 1997; Jones et al, 2001). Such repositories have also been criticised for stripping information of important contextual cues, and abstracting the information (Bonifacio et al, 2002). Unless effort is taken in documenting information in a way that is reusable to different types of users for different purposes, others may have great difficulty in re-using that information (Markus, 2001). Peer-to-Peer systems offer advantages for allowing the user to maintain and use their *own* organisational structures and allowing others access to these. This overcomes some of the use barriers of shared central repositories. However, the issue raised by Markus does not disappear--information may still require work if it is to be re-used by others. If users do not take account of the fact that their information is accessible to others, this raises the question of whether problems of re-use will arise in Peer-to-Peer systems when peers come across information which has not been specially prepared for use by others.

Thus, organising information can be a difficult and time-consuming task whether storing something temporarily or permanently, for oneself or for sharing. Consequently, different tools have been developed to help users organise their personal information spaces. These include "WebBook" and "WebForager" (Card et al, 1996), bookmark organisers (Maarek and Ben Shaul, 1996) and shared information spaces such as "CoWing" (Kanawati and Malek, 2000).

Sharing Information

The methods by which individuals share the resulting knowledge, skills and gathered information from such tasks with others has also been studied. This spans dissemination through people (Mcdonald and Ackerman, 1998; Paepcke, 1996), repositories (Berlin et al, 1993; Markus, 2001), and the Web (Bly, Cook, Bickmore, Churchill and Sullivan, 1998; Jones et al, 2001).

There is an assumption that information from other people is of higher value and quality than, say, that found through a search engine, because it has already been judged in some way by someone else to be of use (e.g. bookmarks, Kanawati and Malek, 2000). Many systems facilitate the sharing of all kinds of information including:

- information gathering and usage processes (e.g, "Footprints", Wexelblat and Maes ,1999);
- documentation know-how (Satoh and Okumura, 1999; "Hyperclip", Sato et al, 2002);
- organisation methods ("Information Farms", Eastgate Systems Inc, 2002; "Mindshare", Van Dyke, 2002; "Grassroots", Kamiya et al, 1996);
- as well as sources, content and products (Keller et al, 1997; "Grassroots", Kamiya et al, 1996; "Information Farms", Eastgate Systems Inc, 2002; "Hyperclip", Sato et al, 2002; Takeda et al, 2000 and "Memex", Chakrabati et al, 2000).

Many of these specifically encourage information to be pooled or collaboratively rated or filtered or judged in some way on the assumption that such ratings or recommendations offer good indications of usefulness to potential re-users (e.g. "Footprints", Wexelblat and Maes, 1999; the "Open Directory Project" at http://dmoz.org/; and the "Open Bookmark Service", Liu, Wang and Feng, 2001). For example, Kazaa.com (an all format peer-to-peer file sharing application) rewards those users who rate the integrity of the files being shared.

However the literature indicates that information does not have an inherent use or value across a general population. 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 re-use information for other purposes (Markus, 2001; Paepcke, 1996; Wexelblat and

Maes, 1999). Additionally, Markus suggests that the purpose of knowledge re-use and whether the re-user is a shared work practitioner, expertise seeking novice or a secondary knowledge miner will affect what a person needs from the information and therefore what that information should look like, greater modification being required when the re-user is not a shared work practitioner. This mirrors Sellen and Harper's (2002) arguments that it often requires work to make information sharable and that a certain amount of shared domain knowledge or history is needed to make sharing effective. 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 et al, 2002; Iamnitchi, Ripeanu and Foster, 2002). Mirroring this has been the design of technologies specifically aimed at the sharing of information between different audiences. For example, "Answer Garden" facilitates the finding and sharing of information from experts to novices (Ackerman, 2002) and systems such as "Jibe" and "Groove" (Jibe Inc, 2001; Groove Networks, 2003) both examples of emerging Peer-to-Peer systems, facilitate the sharing of files within workgroups.

However, the literature warns against making too many assumptions that colleagues who work together actually use the same body of information. McKenzie and Cockburn (2001) found that within the same computer science department, 90% of the pages surfed (16,290 pages in all) during their study were seen by only one person out of seventeen. Only the university homepage had been viewed by all and only 30 pages had been viewed by eight or more people. Thus, although people may be accessing the same websites through, for example, the same organisation, the content of what they look at may have minimal overlap. Similarly Sellen and Harper (2002) point out that it is sometimes wrongly assumed that because people do similar jobs that they work collaboratively. Defining what we mean by a so-called community of interest (a community whose individuals do have a need for a common body of information or who share the same purposes for that information) is therefore an important issue.

Several tools have been aimed at discovering others with shared interests or purposes. For example "K-Media" (Takeda et al, 2000) identifies others who have similar webpages bookmarked and uses this to find others with similar topic interests in order to recommend pages from other people's bookmarks. Similarly, "Jasper II" identifies shared interests and gathers and shares information within communities of interest through the use of personal profiles and agents (Merali and Davies, 2001).

Peer-to-Peer Technologies for Information Sharing

One of the technological underpinnings of some of the ideas within ePerson (both in terms of information seeking and in terms of sharing) is the notion of Peer-to-Peer computing. Enthusiasm surrounding Peer-to-Peer systems has been spurred on by sites such as Napster which allowed and promoted the sharing of music files between individual users and what they kept in their own personal digital file systems. These new decentralised models of computing hold out the promise of opening up the world of information beyond pre-defined networks such as the Web, and reaching into the systems of individual users. While there is some dispute over the proper definition of Peer-to-Peer 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 PCs.

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 documents, multimedia or other kinds of content. Milojicic et al (2002) provide a comprehensive review of the definition and attributes of some of the current Peer-to-Peer systems available.

One aspect of this that interests us is how this new vision is beginning to spark new ideas for ways in which people might share knowledge, and new ideas for tools to help people manage and access this 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 PCs, 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." (Breidenbac, 2001, ppg. 26).*

There are now a few commercial products that allow you to do this. For example, Groove software allows physically dispersed workgroups to work together on the same project using Peer-to-Peer file sharing. While Groove is aimed specifically at the support of small working groups, other possibilities for Peer-to-Peer systems are centred on the formation of new kinds of virtual communities of people having common interests or expertise. For example, software-based agents might seek out other users who have similar profiles or interests, or who own data whose semantic structure matches the structure of one's own data (e.g. Kanawati and Malek,, 2000; Merali and Davies, 2001; Takeda, Matsuzuka and Taniguchi, 2000). Thus there are a host of technological aspects (e.g. the use of personal profiles, search algorithms, agents, alerts, the Semantic Web, meta-data and so on) that could be used for creating virtual communities, forging new relationships, and leveraging all types of information through these new networks, information that was hitherto under the lock and key of individual ownership. Many such sharing tools (e.g. Kanawati and Malek, 2000; Merali and Davies, 2001; Sato, Abe and Kanai's Hyperclip, 2002; Jibe Inc, 2001; Bonifacio, Bouquet and Traverso, 2002) have moved towards using distributed or Peer-to-Peer architectures.

One technology that is seen as "crucial" in the development of Peer-to-Peer file sharing is the use of meta-data or "the Semantic Web". This is because, unlike the sharing of music files, which has a certain predictability and format (e.g. artist, song title or album), the sharing of other documents or files is likely to require more complex ways of querying and searching. For this reason Kazaa users are asked to add their own meta-data, or keywords, to documents they have created themselves to enable the search process to work effectively (see http://www.kazaa.com/us/kreate/index.htm). Therefore projects such as Edutella, SWAP and our own ePerson project here at HP Labs, Bristol (see Nejdl, 2002; SWAP project, 2002; Banks et al 2002) all seek to combine Semantic Web technologies with Peer-to-Peer systems.

Our concern however was to look at these developments from a user perspective and our approach in doing this is outlined below.

Our Approach

The concern of this study was to explore both information gathering and information sharing from a user's perspective, to understand better what it is that new technologies might be able to support, as well as to understand what possible barriers might be revealed by looking at what people really do. We wanted to look at how information is sought through a network, what information is extracted, what is done to this information, how it is combined with other sources of information and where this information ends up. Going back to our opening scenario, we hoped that this might provide insights into how "Sam" might search for and select information, what "Robin's" resulting collection of information might look like, and cast some light on the potential usefulness of this information for someone like Sam.

In order to explore this we decided to focus on a group of people already known to carry out a significant amount of information gathering: knowledge workers. Gathering information, transforming it, learning from it, and communicating it to others are key activities for knowledge workers. They are also interesting in that their work tends not to be routine, but changes on a project by project basis. Knowledge workers thus tend to use many ad hoc ways of gathering and using information, drawing on a diverse set of tools and resources to do so.

In addition to focussing on knowledge workers, our second main focus was on the Web. The Web has now firmly and inextricably taken its place as an important information resource particularly for knowledge workers. Having said that, it is well understood that the Web is one of many sources of information that they may turn to, including personal contacts and personal collections of documents, these along with the Web being used to differing degrees amongst different professions (Turner, 1997). Arnott and Tan (2000) suggest that choice of information source can be based on rational choice (e.g. the physical attributes of media) or on social interaction (e.g. the influence of other users, past behaviour and norms). They further suggest that the value of information is judged according to its source, how current it is perceived to be and how complete it is perceived to be. For example, internal information is judged as more reliable than external. They argue that, for these reasons, and because of a culture of internet usage, the Web is often used as a source of information in the workplace. One set of workers known to utilise the Web in this way are knowledge workers.

In addition to this, our own recent study (Sellen, Murphy and Shaw, 2002) of how knowledge workers use the Web has shown that "information gathering" is the main kind of Web activity that such workers carry out in the course of a working day. Information gathering can be defined as using the Web to purposefully find and collate information around a specific topic or theme. This includes, for example, gathering information:

- in order to compare, choose or decide about something;
- in order to supplement a future task (such as collecting background information to write a document, or to prepare for a meeting);
- or in order to be inspired or get ideas.

Such activities very often involve sets of questions, ill-defined questions, or questions which are formulated in the course of carrying out a task, this process being supported by the information gathering models outlined earlier. Information gathering is very different from some of the other kinds of Web activities knowledge workers do (such as fact finding). These activities are also often time-consuming and complex, involving the navigation of multiple links and sites, and involving scanning or skim-reading large amounts of material to assess its relevance. This initial study has also pointed toward other issues or important characteristics of information gathering when it is carried out using the Web as a main information source. For example:

- One key issue for knowledge workers engaged in this kind of activity is knowing which sites can be trusted to give comprehensive and accurate information. As a result, they tend to return to a small handful of trusted sites (as opposed to doing keyword searches), and spend considerable time assessing the quality of information on an unknown site.
- One of the ways in which knowledge workers assess the trustworthiness of sites is by learning about them from their peers. Thus, leveraging knowledge in their own community of practice is an important feature of their work and one that impacts their use of the Web.
- Another characteristic of these Web activities is that they are often time-consuming, sometimes taking place over days and even weeks. As a result, knowledge workers need to save the interim results of what they have done so that they can return to their activities. Both paper and electronic tools and resources are used to do this in a variety of interesting, idiosyncratic ways, but it is clear that the tools to hand (such as the current history mechanism as well as bookmark folders) offer inadequate support for the kind of knowledge management and archiving they need to do.

Given the characteristics of this kind of activity, and given its importance for knowledge workers, information gathering activities offer up a potentially good focus for studying the lifecycle of information gathered from the Web and to explore issues relevant to ePerson and other similar tools. Areas of overlap include: the sharing of knowledge between peers, including information re-use amongst a community; the use of personal profiles to provide more efficient and effective Web searching; the development of flexible history mechanisms to support work which may have been interrupted or delayed; and the use of agents to conduct autonomous or semi-autonomous searches and knowledge management tasks.

Method

The study took an exploratory approach, aiming to capture a rich amount of data for qualitative analysis as opposed to statistical hypothesis testing.

Participants

We began by selecting 16 different knowledge workers across a diverse range of knowledge work. Participants were recruited via email, in which knowledge workers were defined as people whose paid work involves significant time gathering, finding, analysing, creating, producing or archiving information, "information" being anything from documents to drawings to multi-media files. Participants were also pre-selected to be regular users of the Web for their work tasks (although non-work related information gathering tasks were included in the study when they arose). Regular Web use was defined as use of the Web at least 4 times in a typical working day.

Table 1. Summary description of participants

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

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, this basic information being collected from the participants using a short questionnaire prior to the in-depth interview. Participants were given shopping vouchers for taking part in the study.

Procedure

Each of the 16 participants was visited individually at their place of work in order to take part in a videotaped interview in front of their PC's. Prior to the interview, it was ensured that each participant understood and consented to taking part in the study. A short questionnaire was used to double check eligibility and to collect some background information on the participants. Having been given a definition of information gathering (see below), they were then asked to identify five or six information gathering tasks they carried out using the Web from the past couple of weeks (using their history list if needed), and to identify which of these were completed or ongoing.

Definition of Information Gathering
"purposefully finding and collating information about a specific topic or theme" e.g. in order to compare, choose or decide about something; in order to supplement a future task (background information for a report/ meeting); in order to be inspired / get ideas. This is different to finding a "specific fact" such as a telephone number.

Participants were then asked to verbally "walk-through" at least two of their tasks. In fact, most participants covered far more than two tasks. Each participant was prompted to start off by explaining the task they had done 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 as they "walked through" the task. Participants were 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?
- 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 re-use anything?

Data Collection and Analysis

Data were collected in the form of videotaped interviews. These were transcribed with the addition of notes concerning the objects under discussion (bookmarks, webpages, printed documents, colleagues etc). This material was then analysed task by task for emergent themes concerning where the information had come from, what was extracted and what happened to this information. Users' reasons for their actions were used to help to understand why the information had been gathered and used in the way that it had.

A broad overview of the emerging themes was discussed with the ePerson group at HP Labs. Given the amount of data collected, the aim of this was to identify some of the most pertinent issues to focus on and to analyse in greater depth. This prompted a more detailed set of questions, such as:

- What are the different strategies people employ for finding information on the Web?
- What are some examples of what people do when their strategies fail?
- How do people find out or know about the resources (links) that they use or revisit?
- What examples are there where people may have re-used Web information from a previous task?
- What type of Web information ends up where (printed, hard drive, bookmarks etc)?
- To what extent do users intend to share information as part of the task they are doing (and how often is it post hoc)?
- When sharing takes place, how often is this proactive on the part of the information gatherer and how often has the recipient requested it?
- What are possible barriers to re-use of information?
- What are the possible barriers to sharing?
- What is being shared with whom?
- In what ways do users "add value" to the information they share?
- What are the methods of sharing (e.g. email, publish on Web, conversation, shared folder)?

The findings below aim to provide an overview of the information gathering process but, within this, the goal has been to provide a greater level of detail regarding the issues outlined above. In the first section we cover the information gathering lifecycle before going onto look at aspects of sharing.

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. Issues that pushed participants towards using the Web included the availability and range of information (including being able to get the latest up to date information), the speed and ease of accessing information, the advantages of digital over paper formats (space and organising), low cost and enjoyability. Factors pushing them away from the Web included the lack of certain types of information (especially information only available through people), wanting information in hardcopy format and also ease and speed of getting information.

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 done, or if you will, 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.

Starting points

For most of these tasks, 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[1], although sometimes they used a combination of the two (Figure 1).

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;

> *"I saw their product at Interbuild which is the international building exhibition at the NEC and they were handing out A4 bits of information that had their Web address on it so I had a look at it" (Architect)*

[1] For the purposes of this study Web Directories such as The Open Directory at http://dmoz.org/ are included in the category of Web Search Engines.

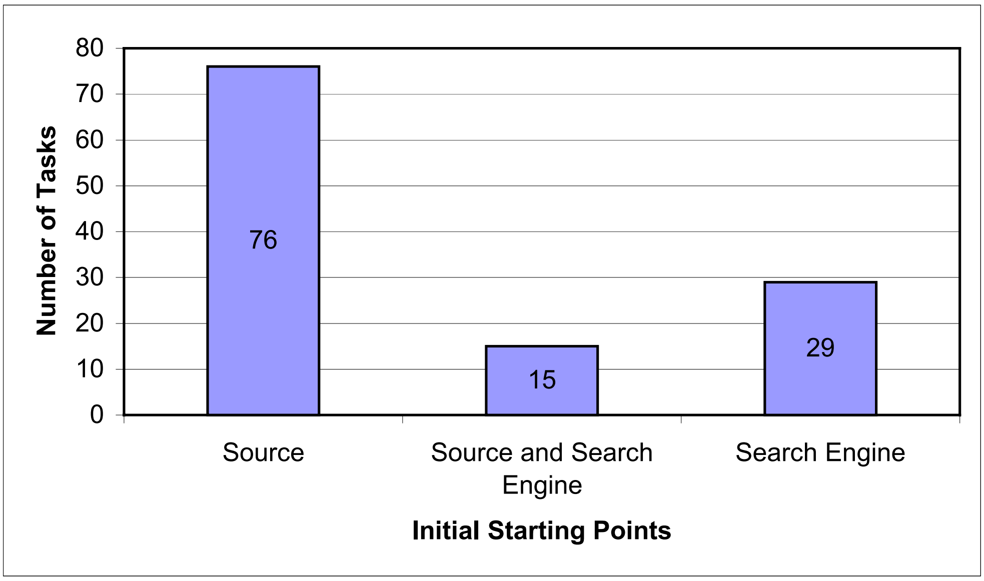

Figure 1. Number of tasks in which known Web sources or search engines were used as starting points

Therefore, each participant continuously comes across new potential sources of information that they can add to their pool of known about sources. It would be reasonable to anticipate that users are not going to rigorously test every discovery they come across for potential usefulness. Some will be used there and then as they are discovered, others may be remembered or kept in case they become potentially useful in the future:

> *"I tend to collect them because people recommend them and say oh this is really good (shows bookmark folder full of search engines with sub-folder of more search engines). And so I go to the page bookmark it, put it in my folder and never look at it again..haha too busy, it takes time to get used to these things." (Researcher)*

For those that have been used, the user will have had an opportunity to filter them for usefulness:

> *"I looked at them all and if they were no use I grouped them into another folder (sub-folder) labelled "no use" so I'd remember later on what I'd done and what had been useful and what hadn't, and the others that were potentially interesting I kept in the er the folder one level up."(Researcher)*

Participants' comments support the suggestion that they had developed some strategies or criteria for distinguishing useful from non-useful discoveries. Table 2 summarises the main issues that emerged from the data.

Table 2. Criteria used for filtering new sources of Web information.

Negative aspects (push to reject)	Positive aspects (push to accept)
Not new, novel or interesting	Provides a lot of information/ gives comprehensive coverage of the topic
Already have a source/ searcher for this type of information	
Doesn't cover the right topic/ domain	
Doubt the accuracy of the information	Information is accurate/ good quality
Not up to date	Information is up to date
Is not easy to use	Information is easy to use/ find
Have seen and rejected this source before	
The presentation is amateurish/ not well laid out	The presentation is "cool"/ well laid out

Thus, there is a process which occurs, be it consciously or not, in which people learn about the information in a source, judge the quality of a website and the quality of the information it delivers. How easy it is to browse through or find the relevant information is also an important factor in determining whether people come back to a source later. Note that some of the assessed attributes will not stay stable over time. The user's task and the information they already have will affect the relevance of a source for future projects and its ability to offer something new.

Perhaps because an effort goes into assessing a site, and judging its usefulness, sites are often revisited. We could tell that at least half of the known sources had been visited before because they were accessed via bookmarks or self-authored webpages. Tools such as bookmarks, webpages and email helped to keep or make certain sources salient and may influence the likelihood of them being used. However such tools can fail if they present sources at the wrong time (e.g. email) or they are not well organised (in bookmarks or on webpages). Other factors such as a source having proved its usefulness or having gained a reputation as "the" source to use in the industry or for a particular task also affected which source from the user's pool of possible known about sources is actually selected.

The knowledge that participants had built up about some of their more often used sources with regard to how and when they could be used to support them in their tasks was obviously quite considerable;

> *"that, for example, tells you all about sql servers but that tends to be big subjects whereas Microsoft tends to be more very specific stuff for me so I find Microsoft doesn't very well explain big stuff sometimes. But it explains something like backups quite well." (Development Manager)*

Because of this kind of knowledge and experience the choice of source often appears "obvious" to the user when working in a familiar domain and they additionally may have built up a selection of back-up sources to provide alternative routes to information if the first source fails;

"Microsoft, their official response is, this is a known bug with the product and will be resolved at a later date or something like that. Not helpful ok, cos you've got 200 people saying I can't download this file or whatever…But there are other places to look on the Web for solutions.. (such as) newsgroups" (Network Analyst)

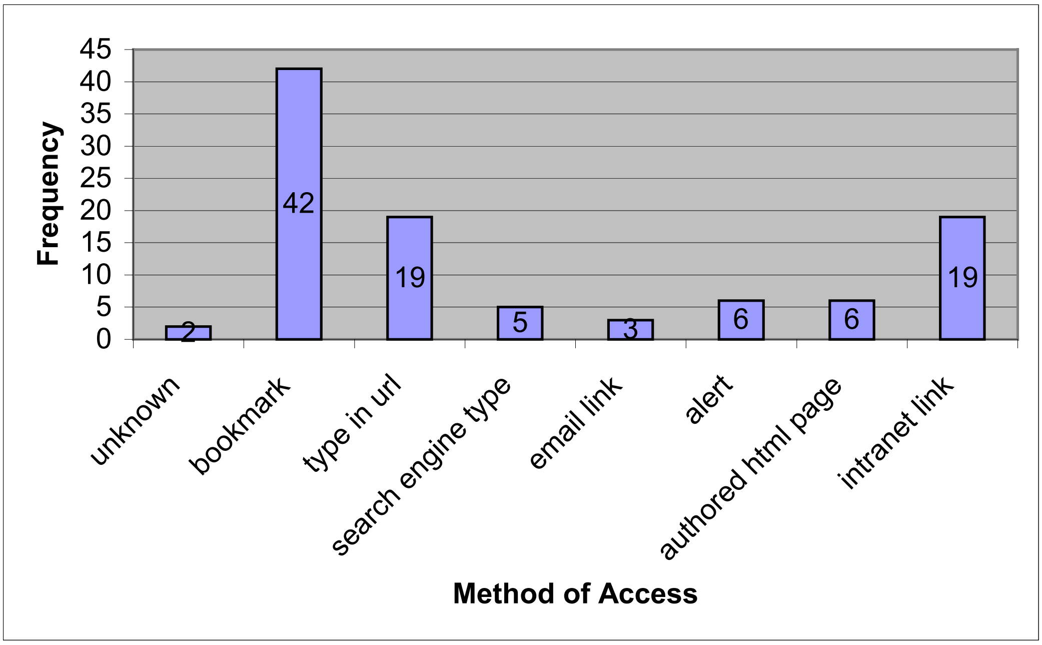

Figure 2. Frequency of methods used to access to known sources.

Having chosen a place to start, participants had a variety of methods for accessing their starting point using bookmarks primarily but also by typing the Url into the address bar, typing into a search engine, clicking on an email link, using alerts, clicking on an authored html page or an organisation's intranet page (often the user's homepage), as is shown in the Figure 2[2].

It was possible to see how some of these methods can be used to access a source directly when only the "name" is known and no tangible Url is stored or saved anywhere:

"here's a (teacher's) Union that I never, I don't think I've ever looked at their website but I'm pretty certain that (guesses Url for the Union website, typing this into the address bar of the browser) will get me there…. , he said showing off (The Union website homepage appears in the browser). Yeh. So that's a fairly, a lot of addresses are you know fairly easy (to guess)" (Union Representative).

[2] The total frequency exceeds the number of tasks that used a known source as a starting point as some tasks had more than one source used as a starting point. E.g. a database and an organisation's website may be accessed as starting points in one task and so are both recorded.

> *" Yeh used Yahoo or some other search engine and just type it in (name of organisation) and it would come up with it quite quickly cos they're pretty well known", (Policy Advisor)*

All of these strategies and the knowledge learned through these processes (of discovery, assessment and remembering of sources and the use of tools in storing and accessing sources) enabled participants to start their searches from known sources as opposed to search engines.

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. Using search engines then helped users scope out information in a new domain to help guide their search;

> *"But saying that, the tools that we have are obviously quite focused on what (the company's) core business research is, so quite often you get a topic that's just going outside that, ... so quite often for their research I would often look on the Web to see what additional resources there are." (Information Research Analyst)*

Participants tended to have a main search engine, the most common being Google. Comments suggested that factors such as speed, good results, being recommended by others, ease of use, lack of non-search related information (e.g. news or advertising) were all influencing factors in choosing a search engine. With time, people also learn how to use a particular search engine and this may make them less inclined to change to another. Having said that, participants resorted to alternative search engines when their attributes outweighed the benefits of using their familiar search engine. This included being able to search several search engines at once (Metacrawler), results being clustered into themes (Vivisimo) which helps to narrow down a search, being able to compare results of different search engines side by side (Opera), being able to search using a question (AskJeeves) and being able to search by category when it is hard to cut out irrelevant material using keyword searches (Open Directory).

What the data show then, is that for most tasks, these knowledge workers more often than not dealt with familiar topics and used familiar resources to begin their information gathering tasks. This is doubtless a reflection of their knowledge and expertise in their particular professional domains. Part of that expertise is knowing where to find information, and how to find it. In terms of gathering Web information, they are likely to re-use sources that have proved their worth in the past, that they feel they can trust, and which they know how to use through experience.

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. Participants not only learned "information" specific to the domain, but they also learnt both general and domain specific Web search skills. 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)*

Learning to focus a search, formulate appropriate keywords and queries, find shortcuts, and identify potentially relevant information, were all skills and knowledge that were developed with experience and built confidence;

> *"The US is really good for environmental work, environmental organisations especially education or environmental education stuff and they tend to post a lot of their um, a lot of their activities….how do I know, that's a really good.. that's experience I think, you know when you do a search and you forget to click that little search in the UK box and say you for example, ... you just type in "education activities" you will find that the majority of the sites you get are American" (Education Officer)*

Thus although participants were gathering task specific information during a search, they may also pick up strategies that will help them to search, identify and gather information in future tasks. This was seen as a key skill for knowledge workers, effecting their ability to carry out their work. Again such skills were seen as a reflection of their experience and expertise.

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 at different points in the search most commonly by copying and pasting into documents, saving whole documents as files, printing, bookmarking, and to a lesser extent by archiving in email, making written notes or saving in personalised Web folders. Figure 3 shows for each method the number of tasks in which the method was used (note that reading is assumed for all tasks).

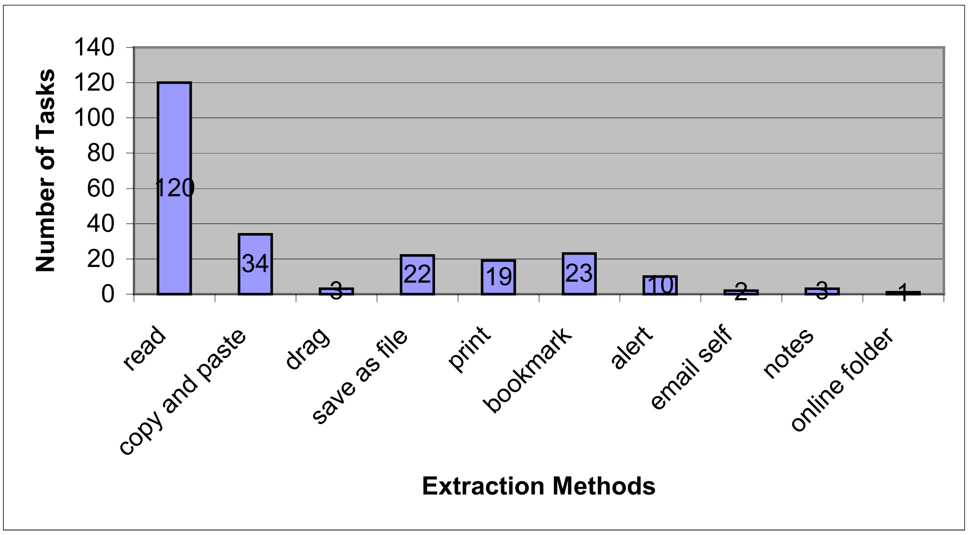

Figure 3. Frequency of use of different methods for extracting information from the Web

Users may also need to follow up some of the information they have extracted by contacting the Web information providers (e.g. company, author, contact name given on the site) themselves (by email or phone) because all of the information they want is not on the website, or it is not in the format wanted, or because it will be quicker and easier to get the information they need through contacting someone. They may also then get information that is personalised to their specific needs.

Generally users did not express problems occurring later because of a failure to take all of the information they needed at the time. Often with the information users had in their head (e.g. names and paths to sites) plus their known sources and the use of history, users were confident that they could find things again if they needed to. Whether this was really true or not, none of the users expressed any concern about problems caused later by not saving something at the time apart from the user who relied on their history (because the history function was not working properly on the organisation’s network at that time).

Indeed, comments suggested that if they later needed more information it was better to retrieve this from the original source rather than from their own personal information archive because then they could access the most up to date information from the Web and also may come across new information that otherwise they would not have known about. Other issues such as expecting the usefulness of the information to be temporary and wishing to keep their archives small and manageable again could deter participants from extracting large amounts of information.

Consequently, any particular information gathering task could have associated with it several different informational “by-products” in both the digital and physical world. Any of these by-products could be reformatted or otherwise modified or transformed more than once. Some of these by-products were shared and some were not.

For example, in looking for ideas for illustrations, the graphic artist bookmarked webpages, printed them off or dragged images from the browser onto the desktop which were later dragged into Photoshop. Images saved in Photoshop format were also sometimes printed. The printed webpages and Photoshop images were collated with photocopied pages from books into a plastic folder which was accessible to work colleagues. Later the Photoshop files were archived onto CD.

As another 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 these examples illustrate, any given information gathering task may have associated with it many different kinds of informational by-products, in many different formats. Each serves a different purpose, some of which are transient or temporary, and others which are useful in and of themselves. Understanding how they are related and where they have come from may be quite complex.

Storing/ Archiving

Looking at a snapshot of where these by-products ended up in participants' own information spaces (Figure 4) also reveals some interesting trends. Figure 4 distinguishes where Urls were stored as against any other kind of extracted Web information. Bookmarks were the most common kind of format, followed by email, html pages and Word documents. Less common were PDF, program, text, image, Photoshop, Excel and CAD files. In only 11 tasks was there no physical extract of information from the task, not even a re-used link stored from a previous task. In all cases this was not because the user had not recorded anything at all, but could be because they had passed the information onto a recipient (e.g. printed sheet) and not left any other trace for themselves.

Figure 4 also shows that information could be stored in places that were either accessible to the user only (personal spaces) or that were also accessible to others (shared spaces). The main personal spaces used were application folders on the hard drive or organisational network, bookmarks or Outlook Express. Shared spaces were mainly intranet pages, shared network folders or webpages. Printed information was either stored in a person's own work area, or made available to others through shared office shelving.

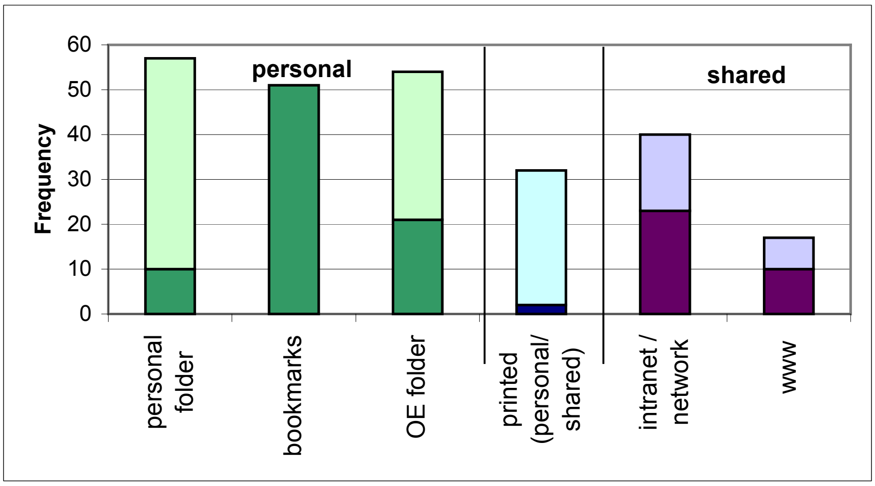

Figure 4. Personal and shared storage places for Urls (shown in dark shades) and other Web information (shown in light shades).

Some of the issues that emerged regarding the nature of personal versus shared spaces are summarised below. These involve access, content and organisation:

- **Access.** Personal spaces were those that were accessed by the user alone, shared spaces were accessible by other people in addition to the user. It is interesting to note that firstly barriers to access may be more social than physical for personal spaces (e.g. other people can access a network folder, hard drive or printed material on a desk, but they don't) and that the user may not be completely aware of exactly who has access to a shared space (e.g. exactly who has access to a shared network folder or whether a html page is on the intranet or the Public Web). But there was no evidence of users being concerned about these issues. Users either trusted their colleagues not to access their personal spaces or that only those who should have access would have access to shared spaces.

- **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. They (personal spaces) can also be used as a dumping ground for information that does not belong anywhere else (e.g. "when I clear my computer I stick stuff in here", (Development Manager)) 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.

- **Organisation.** Although personal spaces may be described by users as fairly organised by topic or project, it was pointed out by one user 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) and also who he was currently working with in order to understand and make use of the information. 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 folders category may be put there and similarly information that could be filed was not. By contrast, shared spaces demanded a more consistent organisation so that others could find information easily. This was particularly of concern to users themselves when others could contribute information as then they also had the task of finding other people's information in that space. Multi-contributor shared network spaces meant things may 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 also come into play in managing shared spaces.

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 re-used sites and sources as starting points, in only 3 of 120 cases was any content or were any documents from past projects re-used. In addition, when asked, participants said they expected to re-use 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, but I'm not saying here that it's a really good structured individual thing because (you) will want to take pieces of it you know, you just want to, its not even a jigsaw, like cooking almost, you take all these relevant bits and you mix them together to make your own recipe"*

Comments suggest that reusability is linked to whether the information has any persistent value. For example, some information discharges its value as soon as it is used, or as soon as the task or project it is associated with, finishes. This doesn't mean that this information is not archived somewhere. Users may trade off costs of space and effort to organise archives with the need for record keeping and keeping information "just in case". However, users' anticipation that they will actually re-use such information can be very low. By contrast, some information will have a persistent value because it is easily modifiable or applicable to similar tasks that are likely to occur in the future, or because the exact same task will repeat itself. Although a minority of these involved information from past projects which may be preserved (e.g. laminated) the majority of reusable information can be characterised as sources of information which may be organised for easy retrieval (e.g. bookmark maintenance and hierarchies).

By contrast, what these knowledge workers were re-using were the skills of gathering information, something they learned from long experience. For example, the Architect described the stages that he would go through for a typical task and the points at which information gathering fed into this. Similarly, 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:

> *"It's 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"*

Patterns of Sharing

Turning from the lifecycle of such tasks, we now look closer at the sharing of information in such tasks since this is a key goal of this study. Whether or not information is gathered with sharing in mind, with whom it is shared and how it is made sharable are issues which, in a sense, cut across the lifecycle we have described.

When we looked at participants' initial intentions, in 71 of the 120 tasks (60%) 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).

Sharing was most often with individuals or small groups, and this was done mainly through person-to-person methods (mostly via email, or face to face, occasionally by fax or memo). 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 5. Comments suggested that the type or number of recipients, efficiency and effectiveness, and organisational policies or work practice, often dictated the methods of sharing that were used:

> *"we're a small team we don't want to sit here and stuff envelopes and whatever. You can just go, get your email list and "send" and can go out to a 100 people in 5 minutes so it's you know, saves you time, saves you money and you use less resources which ticks loads of our boxes" (Education Officer)*

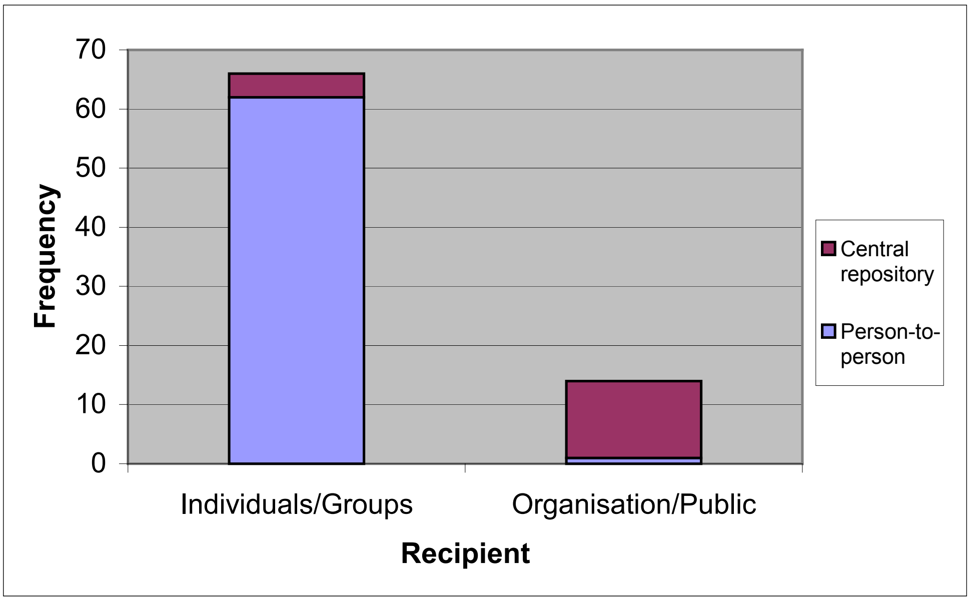

Figure 5. How often information was shared with individuals/ groups versus organisations/general public. Also included is whether information was shared using central repositories or person-to-person methods (e.g. via email)

The method of sharing was also obviously linked to "what" was being shared. Thus, email, shared virtual folders and webpages allowed digital text, images and attached files to be shared (email and shared folders showing a wider range of file types than the Web which tended to be PDF or program files). Post/ fax , memos and shared shelves allowed hardcopy text or images to be shared. Face to face allowed both transient sharing of information that was spoken or shown but not given to the recipient, but it could also allow physical information to be retained by recipients whether that be by handing over a hardcopy printout or information on a CD. We found that in 41% of tasks Urls were included in the information being shared.

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 6). Specifically, we found that shared information could be modified in three ways, none of which was mutually exclusive. It could be:

1. **Rewritten:**
   " So I rewrote , I didn't, I copied the thing but I rewrote in my words (text pasted from Web into Word)… I plonked it all in there and then rewrote on top of it" (Development Manager)
2. **Written"around":**
   "(I've written on there) the person who It's for and that's a record (that he) said he would send the invoice with the report, I just made a note of it on there" (Information Resource Manager)

3. **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, email, faxes or memos at the point of handing it over: *"what I did was I copied and pasted from the website the explanatory paragraph about this research that is being done and I included that in my email to this person asking questions about this research" (Researcher)*

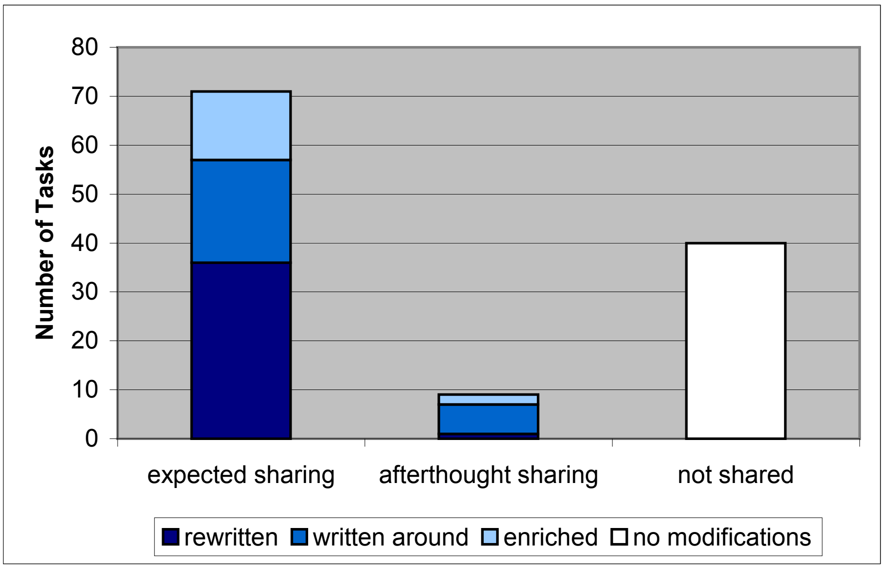

Figure 6. Number of tasks in which people shared information as expected or as an afterthought or in which information was not shared. Also shown is whether the information was rewritten, written around, enriched or unmodified.

The important point to note here is that information that had not been rewritten or written around was at the very least enriched at the point of sharing (distinguishing it from non-shared information). 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;

> *"I just copy and paste the text and put it in a normal Word folder and then I just bung it somewhere and I'll just keep adding things to the same Word folder , things of the same subject so its there if I want to access it in the future" (Researcher)*

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 is concerned that some of his bookmarks reveal his interests and where he banks or shops), or participants didn't know anyone or were unsure as to who may find it useful (e.g. the Researcher copies information into an email to share with work colleagues and then deletes it, not being sure whether it will be useful or interesting to colleagues or not) and additionally not knowing how to share it could influence whether or how something was shared (e.g. the Information Resource Manager wants to share bookmarks but cannot remember how to do this. And similarly, the research scientist would like to share more information on his webpage but this involves the effort of asking someone else to alter the page for him and so this information has never been added).

Making Information Sharable

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 again:

> *"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, harking back to Harvey Sacks in the 1960s (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 webpages, 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.

Customising delivery: At the point of information delivery, we also saw that information gatherers designed and chose how information was shared with others. For example:

- Timing. In some cases, account was taken of when it would be most useful for the recipient to receive the information and therefore to share it at the appropriate time.
- Enriched delivery. Messages, explanation or discussion often took place when information was handed over, 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 (e.g. 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.

Summary and Implications

To summarise, there are several key points to take from these findings, each of which has implications for new gathering and sharing tools:

1. Use of familiar sources

The knowledge workers in this study more often than not used familiar resources and tools in their information gathering tasks. This implies firstly, that knowledge workers have ways of assessing, remembering and reusing sources, secondly that there are advantages to reusing sources (as this was a more frequent strategy to using search engines) and thirdly, that sources are reusable (or “repurposeable”) for new tasks that fall within the same domain.

Sources need to be assessed for whether they are potentially *usable* (i.e. whether the material is of an acceptable quality, standard or content to be used). For example, we found that knowledge workers assessed sites in different ways:

- First, sites were often "scoped out" and assessed for their value taking a variety of factors into account such as its ease of use and perceptions about how accurate, complete and up to date the site is.
- Second, sites will be assessed by their reputation, often through word of mouth by peers, for example. Sites associated with trusted organisations will also be preferred in many cases.
- Third, knowledge workers may have cross-checked a site for its accuracy using other sources (such as books and documents). Thus these become the sites recognised and trusted by a professional community.

For these reasons, a newly discovered source per se is not that useful until the user has some knowledge about its quality, standard or content. On reuse the user has to be able to recall that the source has been checked or is trusted. They must also be able to find the source again. Several strategies support the worker in doing this. Frequency of use allows sources to be easily brought to mind and people also used a diversity of tools (such as bookmarks, emails, history lists, authored webpages and so on) to organise, structure and label sources in idiosyncratic ways that will be personally meaningful to that worker. The data suggested that if the worker was unable to bring to mind a potential, usable source then search engines would be used to discover a new source.

Therefore if a tool only shared links to sources it would provide the user with little more (or even less) than if the user searched for their own source say through Google. The advantages occur when potentially *usable* sources can be identified, which requires knowledge. This knowledge is acquired by knowledge workers allowing them to bypass the effort of searching and assessing sources each time they reuse a source. The implication is that if this knowledge is leveraged for use by others they may similarly benefit and this is the rationale for why we may wish to enable the sharing of sources with new kinds of tools.

Thus there are several implications that follow from this;

- First, given that sources are stored in a variety of places (including a person's head), they will need to be identifiable in order that others can find them. For example, bookmarks contain links for many reasons, they are not all used as sources. Merely providing access to another's bookmarks may mislead the reuser (e.g. some of the most useful and frequently used sources may be stored in a person's head). *Question:* Are there ways of identifying those sources that have been accessed directly (e.g. via bookmarks, the homepage or address bar) or ways of utilising the history record to identify sources used, possibly distinguishing these by different types of tasks?

- Second, if a knowledge worker is to take advantage of someone else's gathered information (from the Web, for example), the more that they can tell "at a glance" about the usability of a source, the less assessment they have to redo for themselves. In particular, having cues as to the quality, standard or content of the sources . It may be that knowing about the expertise of the owner of the source may satisfy the user (e.g. people are seen to use recommendations from colleagues or links provided on internal webpages). *Question*: Are there ways that new tools could offer a "layer" of information about the person that the information comes from? Or about the assessment that has been carried out. This might be a useful feature.
- However, just knowing where information comes from may not be enough for peers viewing other peers' information since they may know little about the reputation of an information source outside their own realm of expertise. *Question*: Are there ways that sharing tools or Semantic Web applications could rank or rate the "trustability" or reputation of sources for different professions or communities? For example, if an information source has been cross-checked for accuracy against another information source, is there some way this information could be extracted and fed back to users? This might be useful information.

2. The ad hoc nature of projects

While we can see that knowledge workers will have gathered a collection of *usable* sources (i.e. the material is of an acceptable quality, standard or content) such attributes do not make information inherently *useful.* We saw that in contrast with the re-use of sources, knowledge workers rarely re-used or re-purposed the end products of their tasks once a project had been finished. Further, they did not really expect to re-use documents in the future. Yet, we would anticipate that such information having come from quality sources shared the same *usable* attributes.

We found that the prediction of future *usefulness* is based upon the content of information in relation to the context of a future task. Knowledge workers gathered materials in a bespoke way on a task by task basis. End-products generally contained material that had been tailored to a specific task and purpose (and is information knowledge workers have often been informed by and consequently carry in their heads, Kidd 1994) whereas sources offered ways of finding and gathering new ad hoc collections of easily accessible and modifiable information to fit the new task or purpose.

Consequently the *usefulness* of information is always task dependent and sources appear to have greater potential *use* for knowledge workers than end-products and are thus more likely to be selected (issues such as sources being more visible and accessible may also reinforce this choice). On reusing sources, knowledge workers have to predict whether they are relevant to the current task and will provide them with information they do not already have. Over time a site proves itself (or not) through experience and so knowledge workers have a sense of what information it can provide. This information is useful in aiding the worker in predicting the usefulness of the source for future purposes. Some sources will be used so frequently for certain purposes that their selection becomes obvious or automatic. The data suggested that if predictions failed (the selected source not providing what was expected) then search engines would be used. (Suggesting perhaps that new tools must support the making of predictions of usefulness with a high likelihood of success if they are to provide the user with benefits over and above search engines).

Another point to note is that knowledge workers may have a larger number of sources for key domains or tasks, firstly, in order to cover the information of interest from different viewpoints or levels of detail and secondly, so that there are back-up resources to hand when preferred sources are not accessible. This again lowers the risk that sources will fail and consequently lowers the likelihood of needing a search engine.

There are at least two implications of this:

- First, the reuser's task is important in predicting the usefulness of information. Given that gathered information is often tailored to a specific task or purpose it may be less likely to match another's purposes than sources are. By implication tools need ways of effectively matching purposes and predicting usefulness. *Question*: Are there ways that tools can match purposes across peers? Alternatively are there ways in which tools can provide cues to the reuser to judge usefulness for him or herself? (For example, users are seen to make use of webpage descriptions on search engine results and use these to predict whether a linked-to page is likely to be useful).
- Second, the knowledge that a reuser already possesses is important in predicting the usefulness of information. *Question:* Are there ways in which the information the reuser already has (or has previously seen and rejected) can be taken into account and use this to only present new sources or new information to the user ?

3. Familiar domains

Knowledge workers not only tend to revisit known sources, by they also tend to stay within their own domains of expertise (after all, this is part of being a professional expert). If this is the case, then one needs to ask a fundamental question about the frequency with which people do seek out information on entirely new domains. We might reasonably expect that some people do this more than others. A challenge for people who develop these kinds of tools is then to be able to pinpoint those user populations who have the highest need for these kinds of tools. This study has shown that the users of these tools may not be knowledge workers or domain experts, although they may be the population on which such tools depend as the providers of information.

4. Learning by doing

The knowledge workers in this study were seen to learn not just from the end products of their information gathering tasks, but crucially from the *process* of searching and gathering itself. We saw that;

- doing searches helped them learn about new domains, to help them find what language and terminology was used, what key players were in an area, etc. They also gained a bigger picture about what was not relevant, or what they were excluding from the picture by focussing in on a particular topic.
- searching is a key skill for knowledge workers. Doing searches not only helped them to gain important searching strategies (e.g. such as "starting wide and focusing in"), they also learnt how to utilise search tools (e.g. how to formulate and phrase a query). Such skills were perceived as highly reusable and essential in effectively gathering usable and useful information. The data suggested that when this information was hard to find people often attributed the lack of information found to a deficit in their own skills as opposed to the inadequacy of search tools or a dearth of quality information on that topic.
- another side benefit to having gone through the process, rather than accessing only the end products is that this then allows gatherers to "backtrack" through material. This may be important because it enables a person to go back and retrieve bits of information that they failed to extract earlier on. Furthermore, finding information from original sources rather than from extracted by-products in a personal archive may have the advantage of being more up to date and alerting the user to the latest issues or news in the area.

The implications are that while peers may benefit from the gathering work of another by helping them focus in on key information, bypassing the process through which that information is gathered may mean:

- Such tools may also cause them to bypass that part of the learning process. For information that is irrelevant, presumably this is not a problem, and is one of the great benefits of the concepts proposed in projects such as ePerson. But it also suggests that there is a great deal of implicit learning that goes on in the gathering that might be undermined. Maybe people can learn as much about a new area as they can from having access to the "end targets", and develop their searching skills, by having access to the navigation patterns of experts, the keywords used, and the search strategies. *Question*: Can tools reflect some aspects of the process by which information was collected in addition to the end products themselves? It also raises the question of whether the metadata in a Semantic Web might provide a useful scaffold from which to build an abstracted picture of a domain. For example, would there be a way of using expert knowledge (as reflected by an expert's search process) to build a "map" of a domain by drawing on the meta-data of different visited sites?

- Such tools may also undermine the ability to backtrack. Because information has been extracted for a specific purpose, allowing someone else to backtrack to the original source may allow them to take the pieces of information they require from this process and repurpose the information for their own task at hand. The fact that knowledge workers often included links to sources in the information they shared suggests that providing a link back to the source is perceived as a useful behaviour. The implication here is that tools should strive to explicitly maintain the link between sources and end-products in order to allow users to backtrack.

A useful point to note is that whereas sources allow breadth-depth searching, end-products allow users to cut straight to the details. Taking account of the background literature in this area, it is suggested that sources may be more useful and comprehensible to novices who need to obtain background information whereas experts may be able to make better use of end-products there being a higher likelihood that they will be able to comprehend what they are looking at and where this sits within a wider picture (e.g. see Footprints, Wexelblat and Maes, 1999). It is not known whether presenting certain aspects of the process, as suggested above, can help to bridge the gap between novices and experts and it is an interesting point to consider especially if an identified user group for such tools is to be expertise seeking novices.

5. The PC as a workbench (not a library)

Looking at what knowledge workers keep on their PCs shows that these information traces (be they documents, images, Web links, email messages and so on) are in fact an amalgam of tools that knowledge workers have gathered or constructed. Just looking at bookmarks, for example, shows that some are gathered so that they can be looked at later (but may not be looked at), some are gathered but then judged not to be of use, and some have been used but are now obsolete. These links may therefore be of little value to the person that keeps them, yet they may take their place alongside or be intermingled with links that are highly valued and used.

Related to this, looking more broadly at the range of documents that knowledge workers use and gather has shown that 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. 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". Documents tended to be poorly organised and to need explanation. For the owners of these documents, however, this was not a problem. For current tasks, they were very well aware of what documents were created for what purpose, which were the unimportant ones, the ones in progress and the final completed drafts. This says that really the way to view a knowledge worker's PC is as a workbench, in which there are a variety of tools lying about supporting work in progress. A finished product may or may not exist on the workbench. This view is in contrast to likening a knowledge worker's PC to an archive of information. Archives may exist within the PC, but they are likely intermingled with the other artefacts of work in progress.

This suggests that one of the things that new tools might usefully do is either to try to provide ways of discerning or extracting some of these contextual details of use from a collection of digital files. For example, one can ask whether there are ways of being able to show how often a file (e.g. document or bookmark) has been accessed; how one document is used in relation to another; whether a document is merely a "waystation" to creating another document; or, whether a document is a final product in itself. This suggests almost applying a document "lens" to a collection of files to enable people looking from the outside in to see a layer of context on top of the documents themselves.

6. The work to make information sharable

Another important finding from this study leads on directly from the point made above. Because many of the documents kept on knowledge workers' PCs need to be interpreted within their context of use, documents that are shared are worked on in order to make them that way. This is done either by modifying the document or information itself, or by adding contextual details to the information. Such details are often added at the point of delivery, such as adding explanations in an email message, having discussions when documents are handed over in person, and adding things such as cover sheets to faxes. Alternatively, the documents are written either with a particular audience in mind who may share already some of the contextual details (such as a colleague working on the same project), or they are "stand alone" documents, which can be more generally shared because they are more or less self-explanatory (such as a document that can be published). The point is that recipient-design is a process which knowledge workers do as a matter of course, whether they are aware of this or not.

One implication of this is that it will not be enough simply to give others access to people's file collections which are an amalgam of the sharable and non-sharable; the valuable documents and the obsolete. Instead, designers of Peer-to-Peer 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. For example, we have seen that documents are often rewritten or modified by doing such things as signposting or highlighting information in order to guide the reader. This happens not only within a document, but across them. Often language may be changed or simplified, or texts reformatted or restructured. We also saw that adding overviews or explanations were important. In other cases, information in a document was maintained and updated to make sure it did not become obsolete. *Question*: Can new tools offer ways in which such things might be done to documents automatically? If not, could they help to initiate contact with the owner of the information so that a dialog could be initiated? Or could they separate the sharable from the non-sharable products?

A point that deserves further investigation are the implications of earlier suggestions to make process information explicit and accessible to reusers. The effects of being able to see the processes and links between documents upon a reuser's understanding of the context and purpose of informational by-products is not known. Currently such links and processes are not visible to an onlooker who is unable to distinguish where the products of any one task can be found or how they relate to one another. It is possible that this type of information, intelligently presented, may support the user in understanding the context of a document and may offer ways of making information sharable. Whether someone is able to make such assumptions from others' documents and exactly what information they need to do so merits further study.

Additionally, the ability to start from an end-product and link "backwards" to sources or "sideways" to related documents begins to make by-products "feel" more like sources (e.g. makes them possible starting points from which to gather related information) an important possible consequence to note when we look at how potentially different this is to the way in which the reusability of informational by-products is currently viewed.

7. Feedback to the sharer

Lastly, we wish to bring the emphasis back to the sharer. Feedback from recipients often provided the motivation for information to be shared and it additionally often led to knowledge workers modifying the content and organisation of the information to better suit the recipients. Additionally, knowledge workers were often not sure exactly who had access to information they had made sharable. Not only does this limit their ability to design for recipients, it also places confidential and personal information at risk. If tools allow the sharing of information from a person's personal file space then sharers need to be aware of exactly which information is accessible to others and to whom.

Conclusion

This study of Web-based information gathering has highlighted the fact that the domain of Peer-to-Peer knowledge sharing is very different from, say, the domain of Peer-to-Peer music sharing. In the case of sharing music, reaching into the personal archives of someone else's music files increases access to a much bigger repository of data, and also allows peers to make contact with the owners of music, including their tastes and preferences, in valuable ways.

In the case of knowledge work, the situation for Peer-to-Peer sharing is quite different. Here, we cannot treat Peer-to-Peer networks as increased access to a vast archive of knowledge work. For what we find is that the contents of a knowledge worker's PC is not so much an archive as it is a workbench. It is a workbench in the sense that it houses the artefacts that support the processes by which information is transformed. Some of that transformation is for personal use, and some is for sharing.

What this also means is that the information that a knowledge worker owns and keeps in a personal system is fundamentally different from the information that is kept on the Web. The Web contains materials upon which work has been done to make it suited for sharing, and for repurposing. In contrast, personally owned information has been gathered, created and organised in ways that are personally meaningful to it's owner and are by-products of the way in which that information has been used and the specific tasks that have been carried out. These environments are not tuned for sharing.

This is not to say that some aspects of personally owned information may not be useful and valuable to others. Indeed this study has provided insights into the prerequisites for reusing information such as;

- the knowledge that people acquire about their information and the benefits that this provides (e.g. being able to effectively find, select and reuse sources)
- how personal information can be made sharable and what is required to make this happen (e.g. having knowledge of the recipient and explicitly using personally acquired knowledge to make the information recipient-designed)

The challenge for new tools is to harness such knowledge and strategies, maintaining personalised information spaces for owners of information whilst providing an accessible and comprehensible view of this to others wishing to reuse their knowledge, skills and information.

References

Abrams, D (1997) Human Factors of Personal Web Information Spaces. [electronic version] *Knowledge Media Design Institute Technical Report #1*, University of Toronto. Retrieved May 23rd 2002 from http://www.dabrams.com/research/bookmarks/thesis/default.htm

Ackerman, M.S. (retrieved 2002, May 1st) *Information and Computer Science, University of California, Irvine.* Available at http://www.eecs.umich.edu/~ackerm/drafts/cscw.cis.overview.pdf

Arnott, D.R., & Tan, W. D. (2000). Managerial information acquisition using the World Wide Web: An exploratory study [electronic version] (*Working Paper No. 1/2000*). Melbourne, Australia: Monash University, School of Information Management & Systems. Retrieved May 16th 2002 from http://www.sims.monash.edu.au/dsslab.nsf/21f101800eeed7abca2569cd001540d6/5CB6606394316C7FCA2569CD003C335D?open

Banks, D., Cayzer, S., Dickinson, I and Reynolds, D (2002) The ePerson Snippet Manager: a semantic web application. *Hewlett-Packard Technical Report HPL-2002-328* . Available at http://lib.hpl.hp.com/techpubs/2002/HPL-2002-328.pdf

Bates, MJ (1989) The design of browsing and berrypicking techniques for the on-line search interface [electronic version]. *Online Review*, 13(5):407—431. Available at http://www.gseis.ucla.edu/faculty/bates/berrypicking.html

Berlin, LM., Jeffries, R ., O'Day, VL., Paepke, A., and Wharton, C (1993). Where Did You Put It? Issues in the Design and Use of a Group memory. *HP Labs Technical Reports, HPL-93-11.* Available at http://www.hpl.hp.com/techreports/93/HPL-93-11.html

Berners-Lee, T., Hendler, J., Lassila, O. (2001, May) The Semantic Web. *Scientific American.* 284 (5): 34-43

Bly, S., Cook, L., Bickmore, T., Churchill, E and Sullivan, JW. (1998) The Rise of Personal Web Pages at Work. *CHI'98.* 313-314

Bonifacio, M., Bouquet, P and Traverso, P (2002). Enabling Distributed Knowledge Management: Managerial and Technological Implications. *Informatique.Informatique.* 1: 23-29. Available at http://www.svifsi.ch/revue/pages/issues/n021/in021Bonifacio.pdf

Breidenbach, S (2001, Jan, 30), Feature: Peer-to-peer Potential. *Network World Fusion.* Available at http://www.nwfusion.com/research/2001/0730feat.html

Card, SK, Robertson, GG, York, W (1996) The WebBook and the Web Forager: An information Workspace for the World-Wide Web. [Electronic version] *Proceedings of CHI '96, ACM Press.* Pgs: 111-117. Retrieved May 1st 2002 from http://www.parc.xerox.com/istl/projects/uir/pubs/pdf/UIR-R-1996-14-Card-CHI96-WebForager.pdf

Card, SK., Pirolli, P, Wege MVD, Morrison, J.B., Reeder, RW, Schraedley, P.K., Boshart, J (2001) Information scent as a driver of Web behavior graphs: results of a protocol analysis method for Web usability [electronic version]. *Chi 2001 31march-5April.* 3(1): 498-505. Retrieved May 21st 2002 from the ACM digital library: http://portal.acm.org/citation.cfm?doid=365024.365331

Chakrabarti, S., Srivastava, S., Subramanyam, M and Tiwari, M (2000) Using Memex to archive and mine community Web Browsing experience. *Computer Networks.* 33: 669-684

Choudhury, V and Sampler, JL (1997) Information specificity and environmental scanning: an economic perspective. *MIS quarterly,* 21 (1), 25-53

Eastgate Systems Inc. (retrieved 2002, March 25th) *Information Farming*. Available at http://www.eastgate.com/squirrel/Farms.html

Ellis, D., Cox, D., and Hall, K. (1993). A comparison of the information seeking patterns of researchers in the physical and social sciences. *Journal of Documentatio*n, **4**9, 356-369.

Groove Networks (2003) Groove Networks. Available at http://www.groove.net

Harper, R. (1998*). Inside the IMF*. London: Academic Press

Hearst, M (1999) User Interfaces and Visualisation [electronic version]. In R. Baeza-Yates and B. Ribeiro-Neto (eds) *Modern Information Retrieval.* Addison Wesley Longman./ ACM Press: New York. Ch 10: 257-322. Retrieved May 3rd 2002 from http://www.sims.berkeley.edu/~hearst/irbook/10/chap10.html

Hoelscher, C., Strube, G (1999) Searching on the Web: two types of expertise. Poster presented at SIGIR '99. Retrieved May 1st 2002 from http://cognition.iig.uni-freiburg.de/members/hoelsch/sigir99-full/SIGIR99-handout2.html

Iamnitchi, A., Ripeanu, M., and Foster, I. (2002) .Locating Data in (Small-World?) Peer-to-Peer Scientific Collaborations. *1st International Workshop on Peer-to-Peer Systems, MIT, March 2002*.

Jibe inc (2001) Jibe Enterprise File Sharing. Available at http://www.jibeinc.com/pdf/JibeFileSharing.pdf

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. and 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 [Electronic version]. *RIAO 2000: Contentbased multi-media access information.* Pp 170-179. Retrieved May 2nd 2002 from http://citeseer.nj.nec.com/414039.html

Keller, RM., Wolfe, SR., Chen, JR., Rabinowitz, JL and Mathe, N. (1997) A Bookmarking Service for Organizing and Sharing URLs. *Proceedings for the 6th International World Wide Web Conference.* Available at http://ic.arc.nasa.gov/ic/projects/aim/papers/www6/paper.html

Kidd, A (1994) The Marks are on the Knowledge Worker.[electronic version] *Human Factors in Computing Systems. Proceedings of CHI '94 Celebrating Independence.* P186-191. ACM: Boston, MA. Retrieved May 2nd 2002 from the ACM digital Library: http://doi.acm.org/10.1145/191666.191740

Knight, C., Munro, M (2001) Visual Information: Amplifying and Foraging. *Visual data exploration and analysis workshop, Proceedings of SPIE 2001, San Jose, USA.*

Kuhlthau, C. C. (1993). A principle of uncertainty for information seeking. *Journal of Documentatio*n, **4**9, 339-355.

Liu, B., Wang, H, Feng A (2001) Applying Information in Open Bookmark Service. *Advances in Engineering Software.* 32:519-525.

Maarek, YS., Ben Shaul, IZ (1996) Automatically Organizing bookmarks per Contents. [electronic version] *Computer Networks and ISDN Systems*, Volume 28, issues 7–11, p. 1321. Retrieved May 2nd 2002 from *http://decWeb.ethz.ch/WWW5/www185/overview.htm*

Malone, TW (1983) How do people organize their desks ? Implications for the design of office information systems. [electronic version] *ACM transactions on Office information systems.* 1(1): 99-112. Retrieved May 22nd 2002 from the ACM digital library: http://portal.acm.org/citation.cfm?doid=357423.357430

Marchionini, G.M (1995) *Information Seeking in Electronic Environments.* Cambridge, Cambridge University Press.

Markus, L (2001).) Toward a Theory of Knowledge Re-use: Types of Knowledge Re-use Situations and factors in Re-use Success *Journal of Management Information Systems.* 18(1), 57-93

McDonald, DW., Ackerman, M.S. (1998) Just talk to me: a field study of expertise location.[Electronic Version] *Proceedings of the 1998 ACM conference on computer Supported Cooperative Work (CSCW '98)pgs 1-10.* Retrieved May 1st 2002 from http://www.ics.uci.edu/~ackerman/pub/98b25/cscw98.expertise.pdf

McKenzie, B and Cockburn, A (2001) An Empirical Analysis of Web Page Revisitation. *Proceedings of the 34th Hawaiian International Conference on System Sciences, HICSS34.* Available at http://www.cosc.canterbury.ac.nz/~andy/papers/hiccsWeb.pdf

Merali, Y., Davies, J (2001) Knowledge capture and utilization in Virtual Communities. [electronic version] *K-CAP '01.*ACM Press. Retrieved May 23rd 2002 from the ACM digital library: http://portal.acm.org/citation.cfm?doid=500737.500754

Milojicic, D.J., Kalogeraki, V., Lukose, R., Nagaraja, K., Pruyne, J., Richard, B., Rollins, S and Xu, Z (2002). Peer-to-Peer Computing. *HP Technical Report HPL-2002-57*, Hewlett-Packard.

Nejdl, W (2002) Semantic Web and Peer-to-peer technologies for distributed learning repositories. *Edutella Reports and Publications, PADLR.* Available online at http://www.learninglab.de/workspace/padlr/index.html

O'Day, V.L. and Jeffries, R (1993) Orienteering in an Information Landscape: How information Seekers get from here to there. *Proceedings of INTERCHI '93, ACM Press.* Pp438-445. Retrieved May 3rd 2002 from the ACM digital library : http://portal.acm.org/citation.cfm?doid=169059.169365

Paepcke, A (1996). Information needs in technical work settings and their implications for the design of computer tools. *CSCW*, 63-92.

Peters RE (2001) Exploring the Design Space for Personal Information Management Tools [electronic version]. *Chi 2001.* Retrieved May 22nd 2002 from http://www.cc.gatech.edu/gvu/ii/persona/Poster_Paper.pdf

Pirolli, P. and Card, S (1995) Information Foraging in Information Access Environments [Electronic Version]. *Chi '95, ACM conference on Human Factors in Software.* ACM: New York. P51-58. Retrieved May 1st 2002 from http://www.acm.org/sigchi/chi95/Electronic/documnts/papers/ppp_bdy.htm

Rucker J., and Polanco, M J., (1997) Siteseer: Personnalized Navigation for the Web [electronic version]. *Communications of the ACM* 40(3), pp. 73-75. 1997.Retrieved May 2nd 2002 from the ACM Digital Library : http://portal.acm.org/citation.cfm?doid=245108.245125 .

Sacks, H. (1992). *Lectures in conversations, Vols 1 and 2.* M. Schegloff, (Ed.). Oxford: Blackwell.

Sato, H., Abe, Y and Kanai, A (2002) Hyperclip: a Tool for gathering and sharing Meta-Data on Users' Activities by using Peer-to-Peer Technology. Available at http://www.cs.rutgers.edu/~shklar/www11/final_submissions/paper12.pdf

Satoh, K and Okumura, A (1999) Documentation Know-how Sharing by automatic process tracking. *Proceedings of the 4th international conference on intelligent user interfaces.* 49-56. ACM Press; New York. Retrieved May 27th 2002 from the ACM digital library: http://portal.acm.org/citation.cfm?doid=291080.291090

Sellen, A, and Harper, R. (2002). *The myth of the paperless office.* Cambridge, MA: MIT Press.

Sellen, AJ., Murphy, R, Shaw KL (2002) How knowledge workers use the Web.[electronic version] *Chi 2002.* ACM Press. Retrieved May 27th 2002 from HP technical reports archive: http://lib.hpl.hp.com/techpubs/2001/HPL-2001-241.pdf

SWAP (2002) SWAP project at http://swap.semanticweb.org/public/index_html.htm

Takashiro, T and Takeda, H (2000) A Context Based Approach to Acquisition and Utilization of Personal Knowledge for WWW Browsing [electronic version]. *Proceedings of the fourth International Conference on Knowledge Based Intelligent Engineering Systems and Allied Technologies (KES2000):* 756-759. Retrieved May 16th 2002 from http://216.239.35.100/search?q=cache:1Pn8Wv4SLR8C:www-kasm.nii.ac.jp/papers/takeda/00/kes00takashiro.pdf+Mindheap+Takashiro+Takeda&hl=en&ie=UTF8

Takeda, H., Matsuzuka, T. and Taniguchi, Y (2000) Discovery of Shared Topics Networks among People – A simple Approach to Find Community Knowledge from WWW Bookmarks. *PRICAI2000.* Available at http://www-kasm.nii.ac.jp/papers/takeda/00/pricai00f.pdf

Tauscher, L and Greenberg, S (1997a) How People revisit Web Pages: empirical findings and implications for the design of history systems. *International Journal of Human-Computer Studies.* 47: 97-137

Tauscher, L and Greenberg, S (1997b) Revisitation Patterns in World Wide Web Navigation. *Proceedings of CHI '97 Human factors in Computing Systems, Atlanta, Georgia.* 399-406.

Turner,K. (1997) Information Seeking, Retrieving, Reading and Storing behaviour of Library-Users. Available from http://citeseer.nj.nec.com

Van Dyke, NW.(retrieved 2002, April 3rd) Mindshare. Available at http://www.neilvandyke.org/mindshare/

Wexelblat, A and Maes, P (1999) Footprints: History-Rich Tools For Information Foraging.[electronic version] *Proceeding of the CHI 99 Conference on Human Factors in Computing Systems: The CHI is the Limit*, 1999, pp. 270 -- 277. Retrieved May 3rd 2002 from http://wex.www.media.mit.edu/people/wex/CHI-99-Footprints.html

Whittaker, S and Hirschberg, J (2001) The character, value and management of personal paper archives [electronic version]. *ACM Transactions on Computer-Human Interaction.* Vol 8 (2) : 150-170. Retrieved May 22nd 2002 from the ACM digital library: http://portal.acm.org/citation.cfm?doid=376929.376932

Wilson, T.D. (1997) Information behaviour: An interdisciplinary perspective. *Information Processing and management.* 33(4): 551-572.