# The Life and Times of Files and Information: A Study of Desktop Provenance

**Carlos Jensen, Heather Lonsdale, Eleanor Wynn^, Jill Cao, Michael Slater, Thomas G. Dietterich**

* School of EECS
Oregon State University
Corvallis, OR 97331
{cjensen, lonsdalh, coach, slater, tgd} @eecs.oregonstate.edu

^ Intel Corporation
5200 NE Elam Young Parkway
Hillsboro, OR 97124
eleanor.wynn@intel.com

## ABSTRACT

In the field of Human-Computer Interaction, provenance refers to the history and genealogy of a document or file. Provenance helps us to understand the evolution and relationships of files; how and when different versions of a document were created, or how different documents in a collection build on each other through copy-paste events. Though methods for tracking provenance and the subsequent use of this meta-data have been proposed and developed into tools, there have been no studies documenting the types and frequency of provenance events in typical computer use. This is knowledge essential for the design of efficient query methods and information displays. We conducted a longitudinal study of knowledge workers at Intel Corporation tracking provenance events in their computer use. We also interviewed knowledge workers to determine the effectiveness of provenance cues for document recall. Our data shows that provenance relationships are common, and provenance cues aid recall.

### Author Keywords

Provenance, Documents, Desktop Search, File Organization

### ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

### General Terms

Human Factors

## INTRODUCTION

Computer work practices have changed dramatically since the days of the first personal computers. Due to the availability of cheap and abundant storage, our personal computers are rapidly becoming black holes for files and digital resources such as: word files, spreadsheets, emails, pictures, and videos. While this in theory means more information is at our fingertips, finding a specific resource is becoming increasingly difficult. Users primarily rely on organizing files into nested folders. Though several desktop search tools like Google Desktop have emerged, computer users still seem to rely on and prefer manual folder search to keyword-based search [1, 2, 25]. As the number of files grow, the number of files in each folder, or the number of folders, or the nesting of folders must increase. At the same time, as the number of files and emails stored on a system increases, the chances of uniquely identifying the desired file through a small set of keywords decreases, rendering these search tools less effective. As a consequence, a number of researchers are exploring the use of provenance to augment pure keyword search by allowing users to specify or see the relationships between files as they search or inspect search results. We refer to these relationships, which describe the history of a file or document's evolution, as provenance. Provenance is a French word meaning "to originate" [20].

There are several ways that files can be related to each other. For example, a PowerPoint presentation written by a group will go through a number of revisions, it will trade hands via email, and contain information pasted from the paper it is based on. In terms of provenance we would see a trail of versions, each building on the previous version. If the group is poorly coordinated (or good at delegating), some branching and later mergers may occur. The presentation will be associated with a number of emails as the authors discuss and send each other drafts, and a number of source documents will be copied and pasted from (see Figure 1 for a simplified example).

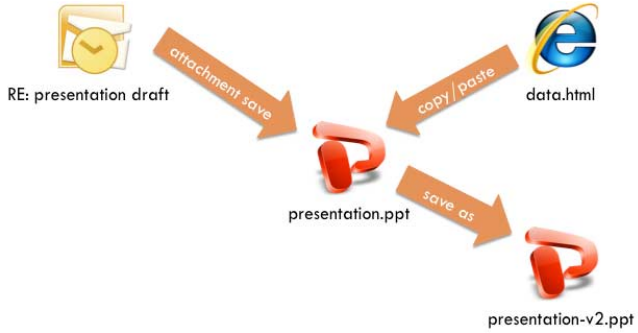


**Figure 1. A sample provenance network.**


*CHI 2010,* April 10–15, 2010, Atlanta, Georgia, USA.

Provenance is important meta-data. While the example above illustrated different types of provenance links, and may offer clues about how provenance might later help a user search for or make sense of the digital artifacts generated. For instance, a user may remember that the document they were searching for was discussed with a particular colleague, or that it some of the content was copied into a specific report. These relationships can be specified as part of a search query.

A user could for instance add a clause stating that they remember emailing Dave about some of the data in the document she is looking for, using that to narrow the list of possible documents. Alternatively, the user may remember that one of the sources for the document was a template the department secretary sent out and use that knowledge to anchor the search. A faculty member may have the final document, but instead be interested in seeing the sources the student used to compare to the list of sources cited, or to determine who the student collaborated or with whom they shared the document. By specifying a provenance relationship, or targeting a specific document, and exploring the provenance links, the user can get a much richer understanding of the document they are reading, or dramatically narrow the possible search space.

The potential of provenance information goes beyond this. Imagine for a moment that one of the authors discovers a mistake in one of the figures. Updating the presentation to reflect the correct numbers might be trivial, but tracking whether and where the mistake has propagated, maybe some of the co-authors pasted the figure on a web site, or cited it in a later paper, is far from trivial. In this sense, the concept of provenance is not new in computing, it is well established in some sub-fields such as database management. There, the goal is tracking the origin of data, and the transformations applied to the data. This is crucial for the later evaluation of the reliability and veracity of said data [7]. This is of course an immensely important ability for those managing databases, especially in instances where these refer to data from different sources. In this paper, we specifically focus on provenance as it pertains to digital resources on a personal computer.

Provenance has only recently caught the attention of the HCI community, and little work has been done to catalogue the types and frequency of provenance relationships in everyday computer use. The primary reason for this is that it requires dynamic event monitoring or observing what the user does on their computer. Current operating systems do not track provenance relationships, so researchers have developed their own tools for monitoring provenance. These tools are promising, but many have not been extensively evaluated. Because these relationships grow over time, longitudinal studies are required to effectively gauge their effectiveness. More importantly for the purposes of this paper, no extensive studies have been done to show that provenance events are indeed common enough to be useful in applications such as search, or to classify the types of provenance relationships seen. In effect, we have not yet seen proof that the approach is technically feasible, and we lack the data to design effective search mechanisms and information displays.

To build the case for the use of provenance on the desktop in general, and document search specifically, we conducted a longitudinal study at Intel Corporation. The rest of this paper describes some of the related work and overview of the challenges we are trying to address – our study methodology, basic statistical results, and an analysis of what these mean for the development of provenance-based tools. We end with a discussion of observed work-practices and opportunities for streamlining the daily life of knowledge workers.

## RELATED WORK

Knowledge workers are highly influential in our modern economy. They are the individuals who perform research, gather and analyze data, and interpret all of the above in order to make decisions and design products. They add value through the information they absorb and possess, and their ability to apply it and develop new understanding [12, 19]. While 20th century productivity relied on manual labor, the 21st century largely depends on the success of knowledge workers. Peter Drucker, who coined the term "knowledge worker", argued: "*the most valuable asset of a 21st-century institution (whether business or non-business) will be its knowledge workers and their productivity*" [14].

The common definition of knowledge work includes individuals working in a great number of different jobs and activities which most information technology (IT) positions fall into, including managers, programmers, analysts, and accountants [12]. Other fields such as design, advertising, marketing, and law, are also examples of knowledge work [19]. Using this criterion, we estimate that knowledge workers currently make up as much as 43% of the U.S. workforce [5].

Knowledge workers typically have a lot of latitude in terms of how they do their work and their productivity and work quality depends in great part on their ability to manage their resources, time, and attention. Their daily routine is one that many of us recognize: A typical knowledge worker spends about three hours a day on "deskwork", two hours in meetings, and two hours communicating with co-workers via phone calls, email and other channels [15, 21]. These tasks are seldom performed in blocks, with workers multi-tasking, switching between tasks and dealing with interruptions. On average, knowledge workers experience 25 interruptions a day (every 11 minutes of deskwork on average) coming from sources such as: email notifications, writing email, making a phone call and hosting in-person visits [15]. These interruptions can have a negative impact on productivity, but tend to more severely affect long-term tasks than to short-term tasks [11].

Productivity can be severely affected when a knowledge

worker has to search for files or information, which is commonly the case with a context switch [19]. Knowledge workers also keep and use large repositories of files and data on their computers and networks. Though their file organization is dynamic, it changes at a slower pace than either the rate at which documents are added to their systems [4], or the rate at which storage capacity increases. Users maintain and archive files at milestones (such project completion), and keep a working set of files spanning back somewhere between six months to eight years [21]. Extensive file and folder maintenance only occurs at events spaced in years, like beginning a new job [4] or moving to a new computer. Classifying documents is a time-consuming task that requires significant effort [3,19].

Researchers have looked for ways to help users deal with growing file repositories. One of the avenues of research has been based to augment keyword search with provenance meta-data. Provenance is traditionally defined as "the history of ownership of a valued object" [20]. In terms of documents, we consider not only ownership, but also the operations performed on the document, especially those that relate it to other people, events, or documents.

One reason for this interest in provenance is that provenance information can be an effective memory cue [3], and may therefore be useful in augmenting desktop search. In order to use provenance effectively, we need to identify effective memory cues [26]. Research shows that although people can recall about 18 computing events within 24 hours, they only remember about four of them one month later [10]. It is therefore essential to track the right types of provenance events in order to aid search.

The biggest hurdle for provenance researchers has been to developing methods for collecting and storing the vast amounts of provenance data generated. The main hurdle lies in the sheer volume and diversity of applications and system components that have to be modified in order to enable provenance tracking. The Provenance-Aware Storage Systems (PASS), a research group at Harvard University, has been among the most active in developing systems to collect provenance data. The University of Pennsylvania also hosted a workshop on provenance in 2007, focusing on the technical challenges of provenance collection and possible uses of provenance data. [7].

A number of researchers have developed systems that use provenance data. These tools typically fall into one of two categories. They either use provenance as search query input or as a way to make it easier for users to discriminate between query results.

Feldspar is a search tool that allows users to incrementally specify attributes and relationships of the file they are looking for [6]. It allows users to search for a document by specifying related emails, people, folders, files, web pages, dates, and events. Users can enter these parameters one at a time, and the list of results updates as each parameter is entered. Feldspar only identifies associations based on static metadata such as email senders and recipients, event organizers and attendees, and received or sent email attachments. Dynamic provenance relationships such as copy/paste and save as are difficult to capture without active system monitoring, and are unsupported in Feldspar.

Another approach to provenance-based search is to adopt a more narrative model [16]. Quill presents users with fill-in-the-blank sentences to prompt users to specify memorable attributes about the document they were searching for [17]. To match the users' answer to files, it monitors system events and stores meta-data about the user's documents, email attachments, web pages, applications, and calendar. Although Quill does system monitoring, it does not capture copy/paste, save as, and other file system operations.

Other search tools use standard keyword-based input, but use provenance to reorder or expand search results. Connections uses traditional keyword search to identify possible files, then uses temporal locality and context to add additional, related files to the results [22]. It monitors file open, read, and write events to detect file input and output. A later version took a different approach by detecting causality relationships. They determined that users perceived results ranked on causality to be better than locality-based ranking [23]. Beagle++ uses email, web, and document metadata to rank search results based on the number of semantic links, which are similar to provenance relationships [8].

Our own TaskTracer system, designed to help users deal with context switching and task management, tracks provenance events on the Microsoft Windows platform [13]. Task Tracer intercepts file manipulation operations in Windows, and has been instrumented to plug into the Microsoft Office suite. This includes save-as, move, and rename operations, as well as copying and pasting within and between applications, and tracking email attachments. It has also been augmented to provide a graphical representation of how different files are related to each other using provenance meta-data [24].

These tools have all shown promise, but are not built on a thorough understanding of the types of relationships that commonly occur between files. This of course has a deep impact not only on the kinds of relationships and links users can reasonably be expected to use in queries, but also the value that these add in terms of narrowing the search space. If a type of relationship is too common, it will hardly narrow the search space. If a relationship is extremely rare, it may prove highly memorable, but may not be worth including in a top-level search interface.

## METHODOLOGY

The goal of our study was to determine the prevalence and characteristics of information reuse and provenance among knowledge workers own repositories (as opposed to those of a workgroup). This required us to track the document and information use of knowledge workers over an

extended period, as we needed to allow time for complex patterns and long-term reuse to emerge.

We recruited 24 knowledge workers from Intel Corporation for the study. Participants were compensated with a $10 gift voucher. 10 of the participants were women (41.67%), and all but one participant was within the 30-59 age-range. 12 participants were managers, four were system analysts, two were administrative assistants, and we had one of each of the following; enterprise architect, software architect, human factors engineer, and senior administrative associate.

Participants completed a survey about their work practices and work style. We installed a custom activity-logging system on their computer designed to record one of the most extensive sets of provenance and information access events to date—document access, moving, saving, copying, pasting, document focus and switching, attachments to emails or web pages. We used a UI-less version of TaskTracer [13] for this study. This means that the task-management and inference functions were removed, as were all the UI elements, leaving only the event tracking system active. We focused on instrumenting Microsoft's Office suite, since this was the most common set of tools used at our research site. We did not incorporate other applications (apart from Internet Explorer), because they accounted for a small portion of overall use, and each additional application required additional implementation and testing by the development team. Participants had the option of turning the system on and off to preserve their privacy and the data was cleansed of all personal and sensitive information before analysis.

We collected at least a month of continuous data from each participant (an average of 43 workdays (8.6 weeks) with a maximum of 63 workdays (12.6 weeks)). Over the course of the experiment seven participants dropped out; four quit due to software/hardware compatibility problems, two had to be excluded because of database problems, and one left on sabbatical.

Partway through the study, we collected data from each participant to prepare our exit interviews. From their database we selected two interesting provenance graphs from each participant's work history (see Figure 2 for an example) and asked participants to tell us a story about the documents, similar to the procedure used in [17]. For each graph, we had a "free-recall" phase where we presented users with document names one at a time, without file extensions, starting with documents more central to the graph and working out to peripheral or "leaf" documents. Afterwards we had a "cued-recall" phase where the same documents were shown with their provenance graphs and participants were asked specific questions about document features, to see if participants demonstrated greater recall. These interviews were conducted either in person or over the phone using screen-sharing software. Out of 17 participants, two could not be interviewed due to scheduling conflicts.

We observed nine of the participants in their workplace in order to document the context of their activities and identify sources of information flow missed by our system, or external to the computer. Each participant was observed for about two hours, during which the researcher sat silently in the participant's workspace and recorded events. This is similar to what has been done in other studies of knowledge workers [15].

We used an open coding method to analyze the interview and observation logs. This methodology helped us identify patterns and common themes in the logs and is discussed in greater detail in [9]. Two researchers coded the logs individually, and a third coder resolved disagreements. Our overall inter-coder agreement was 84.18% based on the Jaccard index (intersection divided by union, or number of codes agreed upon over the total number of codes recorded) [9, 18].

Finally we collected the data and uninstalled our software. Logs were cleaned and anonymized.

## RESULTS

### Overview

Participants worked with a large number of "resources" over the course of the study, where "resources" were defined as spreadsheets, presentations, documents, and other text files, as well as emails and web pages (Table 1). Note that Intel Corporation, where our study took place, like many major companies, hosts a large number of internal applications and resources within a web framework on an intranet. Many of the web resources accessed by our participants belong to this group.

| | Unique resources | Resources/Person-day |
|---|---|---|
| Web* | 65,741 | 89.9 |
| Email | 53,875 | 73.7 |
| Word | 3,208 | 4.4 |
| Excel | 1,854 | 2.5 |
| PowerPoint | 1,555 | 2.1 |
| Text | 275 | 0.4 |
| PDF | 112 | 0.2 |
| **Total** | **126,620** | **173.2** |

**Table 1: Average workload and resource use**

The vast majority of resources used by participants fell into the more ephemeral categories of emails and web pages (94.5% of all unique resources). We refer to these as ephemeral because by their nature they see little reuse. While some emails and web pages are saved for future reference, web pages and emails are generally read, reacted to, and then filed, deleted, or abandoned. This of course does not mean that they cannot, or are not, major sources of information and provenance. Documents, spreadsheets, and presentations made up 7,004 resources (5.5% of all), or 9.6 unique resources per person-day. 299 (4.3%) of these were downloaded from email, web repositories, or shared folders. Figures 3 and 4 show the breakdown of these resources.

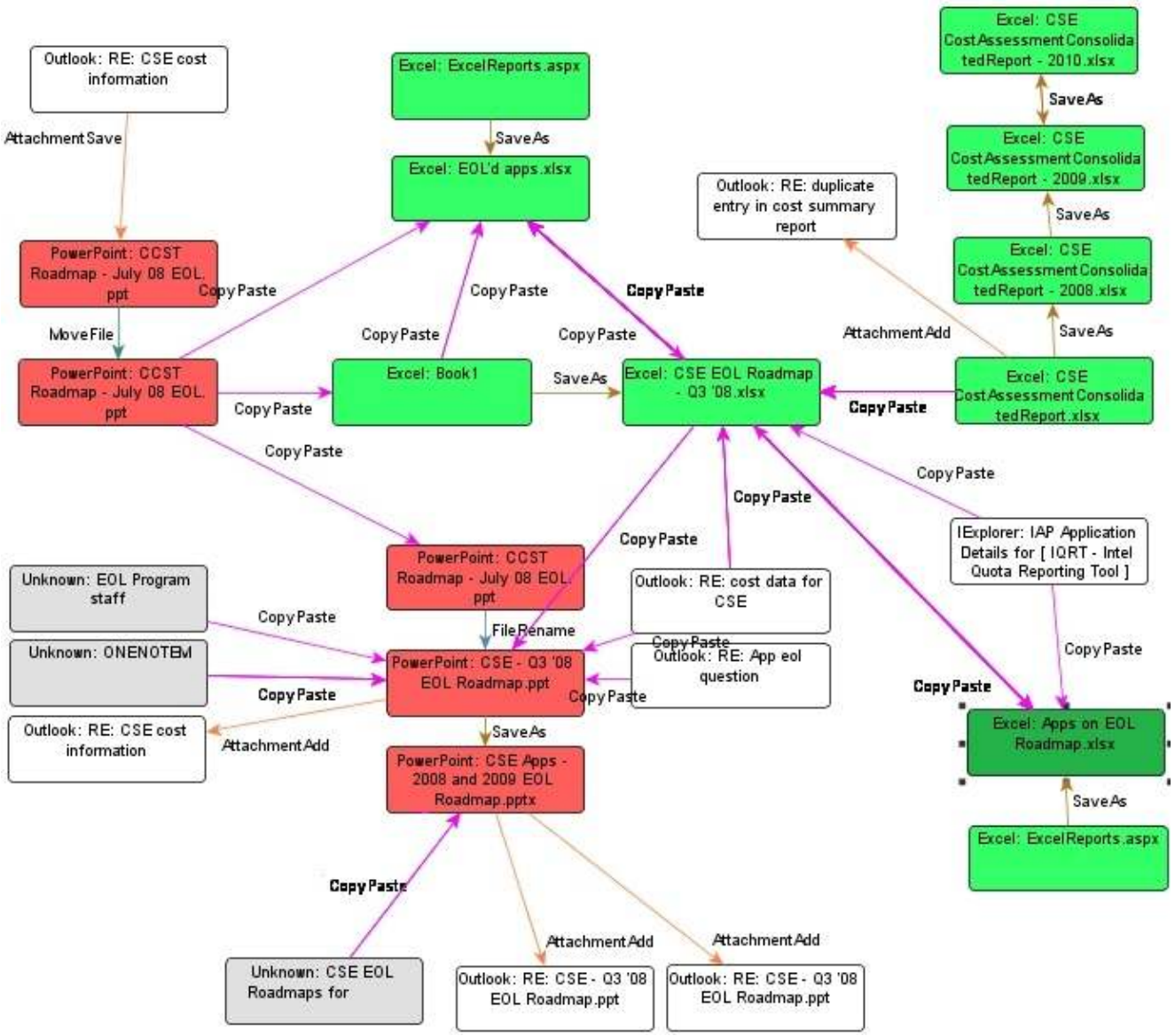


**Figure 2: Sample provenance graph for complex task involving spreadsheets (green), emails (white), and PPT (red).**

| | FROM | | | | | | | TO | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PDF | Excel | IExplorer | Outlook | PowerPoint | Text | Word | PDF | Excel | PowerPoint | Text | Word | Total |
| **Move File** | 20 | 53 | | | 46 | 25 | 119 | 20 | 53 | 46 | 25 | 119 | 263* |
| **Save As** | 0 | 250 | | | 216 | 8 | 231 | 0 | 250 | 216 | 8 | 231 | 705* |
| **File Rename** | 4 | 81 | | | 48 | 43 | 82 | 4 | 81 | 51 | 43 | 79 | 258* |
| **Copy Paste** | 11 | 739 | 851 | 608 | 274 | 34 | 510 | 9 | 983 | 451 | 33 | 1552 | 3028* |
| **Attachment Add** | 4 | 59 | 0 | 0 | 33 | 2 | 43 | | | | | | 141 |
| **Attachment Save** | | | | | | | | 4 | 68 | 66 | 0 | 17 | 155 |
| **Web Upload** | | | | | | | | 8 | 26 | 66 | 2 | 15 | 117 |
| **Web Download** | | | | | | | | 22 | 30 | 47 | 16 | 27 | 142 |
| **Total** | 39 | 1182 | 851 | 608 | 619 | 112 | 985 | 67 | 1491 | 945 | 127 | 2040 | 4809* |

**Table 2. Event counts by application and relationship. (Total =/= Sum of the column as "FROM" and "TO" can overlap).**

Resources had very different lifecycles. The average Word document was open for 105 min. over 1.94 sessions spanning 1.27 working days (not necessarily continuous), while an Excel document was open for 101 min. over 2.11 sessions spanning 1.64 days. PowerPoint was the least used document type, open for 60 min. over 1.76 sessions spanning 1.4 days. While this appears to show that documents led relatively short lives, we did see 705 "Save As" operations (the source of more than 10% of all documents). The user was either versioning a file, or using a file as a template. In other words, files saw far more reuse than one would assume at first glance.

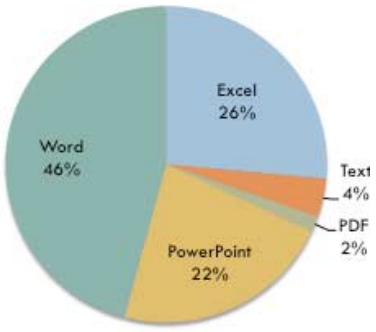


**Figure 3: Breakdown of archival resources by file format**

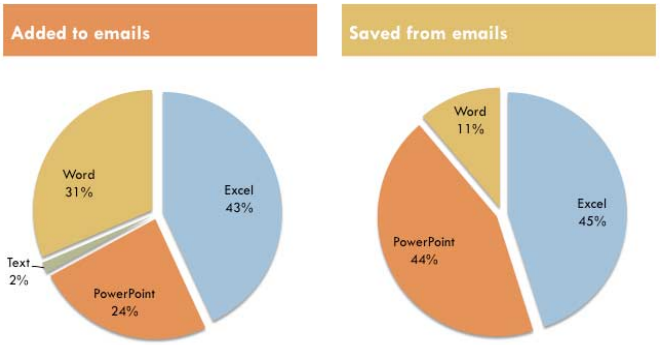


**Figure 4: Distribution of file types distributed through email.**

## Provenance

53.7% of all Excel, PowerPoint, Word, PDF and text files our participants interacted with were related to at least one other file in their personal repositories through a provenance relationship. Our study was not designed to track provenance within groups (though some of this was captured when files were shared via email) due to the logistic and technical difficulties this would entail. Given that all participants were involved in multiple team efforts with multiple different teams, and that we only captured relationships that started after our study started (we were not able to capture or recreate any provenance links that existed between documents already in their personal repositories), it is very likely that we are only capturing a fraction of all relationships.

As Table 2 shows, we recorded a great number of provenance events over the course of this study, over 4,800. While many of the clusters of resources related through provenance events we found were small, involving only 2-3 resources, some of these provenance networks were relatively large (see Figure 2 for an example). We focused our provenance graph analysis on graphs that consisted of at least three resources, which we refer to as significant provenance graphs (simple two-resource relationships are not good examples of provenance history). We discovered 521 significant graphs among our participants (30.6 per subject), with an average of 5.8 resources per graph.

Resources that belong to significant graphs may be good candidates for provenance-based retrieval cues, as they are connected to two or more documents that are likely related. For clarity, we will refer to these resources as "significant resources". Subjects had an average of 178 significant resources in their collections, or four new significant resources per workday. One would expect a much larger number of resources to join these significant networks over time. Significant graphs are important because they are the ones where provenance information will be most useful.

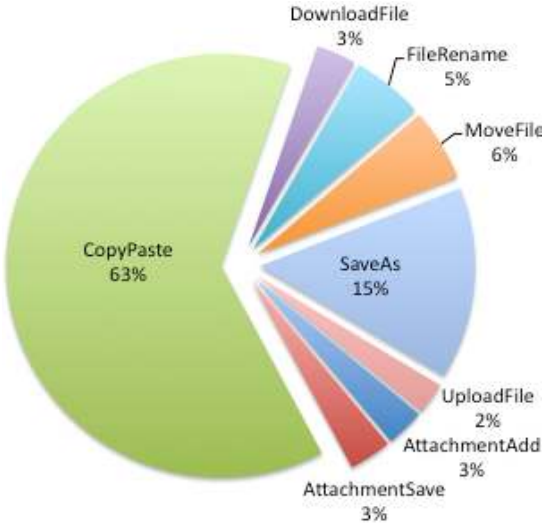


**Figure 5: Distribution of provenance events by type.**

Figure 5 shows a breakdown of the types of provenance events found in our study. When compared them to what our subjects claimed to recall in our interviews and questionnaires about their documents (see Figure 6), we see that there is a relatively good match between the two. The types of provenance events our participants claimed to remember about files in their collections occurred frequently enough to be useful.

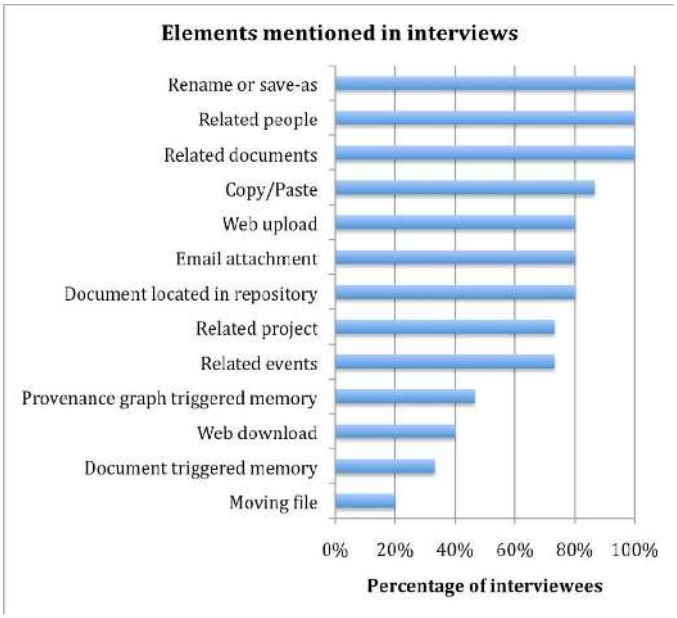


**Figure 6: Common interview elements and percentage of interviewees who mentioned them.**

Copy-and-paste, accounting for 63% of the provenance events we tracked, was mentioned in 87% of the interviews as being important and memorable. The most commonly mentioned "channels" for copy/paste events were through email, websites, and screen captures. File renaming (whether through "Save As" or in the file system) was the second most common event and every interviewee mentioned renaming a file during his or her interview. Though the rest of the provenance events were fairly evenly distributed, email attachments probably occurred much more often than our data suggests, since we could not track all ways of adding attachments to Outlook. 73% of those interviewed mentioned sending one or more of documents through email and 80% recalled receiving a document via email. These instances were often accompanied with comments of copying/pasting and uploading to the web.

When subjects mentioned the location of their file, only 28.6% of the locations mentioned were on the participant's computer (including the user's desktop). The majority of document locations (71.4%) were repositories, primarily SharePoint. This resulted in many web-based upload and download provenance events.

According to our interviews, provenance graphs were effective memory cues, helping users recall more information about their documents. Though this needs to be tested in more detail, the results were promising. Our subjects also understood what the graphs meant. Participants were able to recognize the constructs and reason based on these:

*"It looks like it comes from the IAP tool, and all the green boxes are my Excel spreadsheets that I exported to."*

In many cases, the provenance graph helped users recognize patterns in their workflow. For example, one subject could recall very little about the files until we showed her the provenance graph:

*"Oh, I see… I probably took screenshots and then put them into the document. Yeah, that's exactly what I did. Oh, and I also took screenshots from the document that's not mine. They had wireframes of the B2B portal that I put into my document. Oh, I see what I—yeah… So what I did with the original document was the HFE requirements report out, and probably saved that doc because I wanted to save only a few slides and delete the rest, and probably carried that over to the external portal document… ok and then I carried over to ICSS, yeah. That makes total sense, yeah."*

We also discovered that different types of resources invite different types of use, and hence provenance patterns. Excel is a major hub for cut and past activity. 82.41% of cross-application provenance either originated from or ended up in Excel, compared to 70.11% for Word and only 47.48% for PowerPoint. While these numbers are affected by many rapid focal changes as users tabbed through open windows, Excel was the application that saw the largest number of focus switches (staying in focus on average for 30 seconds at a time), while Word on averages stayed in focus for 3 ½ minutes at a time and PowerPoint on average stayed in focus for 1 minute at a time.

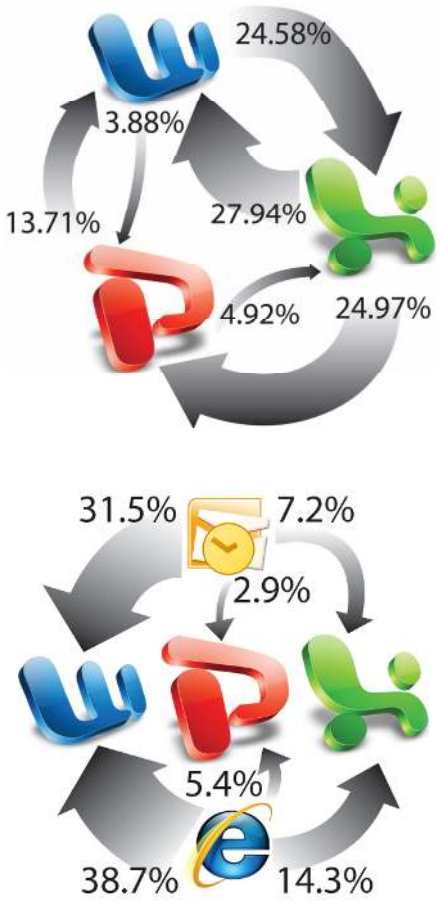


**Figure 7: (a) Information flow across Office applications, and (b) from email and web to Office applications**

Our data seems to indicate that Excel serves as a hub for organizing and processing information, which is then used and archived in Word or PowerPoint documents. Another striking thing is how little information flows out of PowerPoint, indicating that it is largely used as an archival format, much more so than Word and Excel documents. While this may be idiosyncratic to the culture at Intel, the findings resonate with what we have observed in other user populations. Figure 8 has more details of these information flows.

## DISCUSSION

This study is the first of its kind in terms of documenting provenance relationships in actual computer use for several users and several weeks of data. By demonstrating the rich data and relationships we can obtain through provenance tracking, we have shown how the various provenance tools being developed have the potential to impact everyday computing experiences. We have also shown that copying and pasting may be the most common and most memorable provenance link, though most provenance-based search

tools miss these relationships since they track static metadata but not dynamic events.

It is important to remember that the results we present on memorability are all subjective and based on survey answers. It will be necessary to test whether cut and paste events are really as memorable and useful as subjects seemed to think. We do have some concerns about the usefulness of cut and paste events, given how frequent they are. While it will be easy to find events based on these links, they may be of limited usefulness for narrowing search results.

Our findings seem to support the ad-hoc design decisions made for TaskTrail, a provenance graph visualization tool for TaskTracer [13, 24]. A significant benefit of TaskTrail over other provenance-based search tools is that it shows users the context of their results. In our interviews, subjects who did not initially recognize a document in their set found it helpful to see the overall provenance graph. In addition to being an effective cue for re-finding resources, provenance can also help people understand the context of their work and identify related resources. Knowledge workers on average manage about 10 different projects each day and spend a significant time simply managing their tasks. A tool that can show them how they worked with their resources can significantly help in their multitasking and interruption recovery.

Understanding the context of search results is also a significant reason why navigating through directories is still the preferred file search method over search tools [1, 2, 25]. For this reason, we propose that in order for provenance-based search to be successful, it must be integrated in the operating system and traditional folder/directory views. TaskTrail takes a first step by allowing users to right-click on any file in Windows Explorer or Outlook. However, this requires a user to know that they are looking for another file that is related through provenance, much like other provenance-based search tools. Seeing a file in its provenance context helps users distinguish between similarly named files that result from provenance relationships such as copying, renaming, or sending and saving email attachment.

The particular subjects and the particular work environment we studied naturally influence some of the data and patterns seen in this study. Our subjects were picked from a wide variety of groups, with a variety of different job descriptions, but each company and country develops its own work culture, and ways of managing and sharing information. We therefore feel that the specific numbers presented in this paper can be taken less seriously, and that the reader instead can focus on what we attempted to show in this study: Provenance can and does add a rich layer to the study of files and the work patterns of knowledge workers.

As we have mentioned in our results section, there were a number of important limitations to our study. Time was definitely a limitation. Although our study ran for a reasonably long time (8.6 weeks on average), this is short compared to the life of knowledge workers working document and knowledge set. It is reasonable to expect that more provenance relationships would be added to existing files since participants worked on new projects requiring them to weave together new information. This is certainly one of the classic problems in many observational studies such as this: when do you have enough data to make your case? In this case, our goal was to prove whether or not provenance relationships were prevalent enough to be useful for applications such as desktop search, and to come up with a classification of the types of provenance events most commonly seen. Running the study for a longer period would certainly have given us more data, and would have in all likelihood found more relationships, but it is unlikely that it would have changed the distribution of these events at this point. We are satisfied that these events are plentiful.

We encountered a number of limitations when it came to tracking information flow to and from Internet Explorer and Outlook. While the other Office applications had nicely implemented API's that allowed us to capture provenance events reliably, Internet Explorer 6.0 and Outlook proved to contain too many holes and alternative ways of copying, pasting, and uploading and downloading files for us to reliably capture. This was a serious disappointment, as we believe these to be rich sources of provenance.

Early on when designing this study, we decided to focus on the work of individuals rather than workgroups. This was in part a practical decision; it is easier getting the buy-in of a set of scattered individuals than it is to recruit an entire work-group (other than the researchers own). Furthermore, concentrating on a workgroup would have forced us to limit the number of different job descriptions and sub-organizations that were sampled. We stand behind our decision, we believe that this is a better study with more generally applicable results, though we hope to do a workgroup study in the near future to see examine the types of patterns we will discover there.

Much of the focus of this paper has been on how provenance may be helpful or useful for augmenting keyword-based search on the desktop, recording and using provenance meta-data opens up a great number of potentially interesting opportunities for users. Tracking the spread and use of information is a very appealing one for instance. In the introduction we gave an example of an instance where tracking where certain bits of information from a document might be useful. Being able to track down where preliminary budget numbers have been used by tracking the provenance links from the preliminary report is an appealing example. Avoiding accidental plagiarism is the inverse of the same. Being able to assess the reliability of a report as a third party by assessing the reliability of the source information (assuming it is not explicitly referenced). Or allowing a manager to look for breakdowns in communication or collaboration, as evident by a lack of

sharing and cross-referencing of work. These are all relatively simple and compelling reasons for tracking provenance information.

Tracking provenance information implies a certain amount of overhead, both in terms of storage and in terms of processing. The storage overhead increases linearly, as each link simply adds one more record indicating a new link between the two files. The processing penalty is more complex, but by and large occurs when trying to make an inference based on provenance data. In other words, as the provenance record grows, querying the database, or rendering the resulting graphs, grows increasingly difficult. As an example, the graphs used in our study had to be rendered offline, as we did not have a sufficiently efficient data structure, or rendering framework. However, we are fairly confident that these operations can be optimized to the point where it is possible to render these graphs in real-time.

Finally, implementing provenance aware search and file navigation systems may enable us to extend the current file-management model of folder organization. By using provenance data to disambiguate between different versions of files and visualizing their relationships, we may be able to better support more loose and ad-hoc filing behaviors emerging with larger file repositories.

## CONCLUSIONS

Knowledge workers are a critical component of our economy and face many challenges in their work. Between multitasking and multi-teaming, interruptions and information overload, the work practices inherent to knowledge work need sufficient support in order for workers to manage their work effectively—even if it is "all in their heads". The bulk of knowledge work involves gathering, analyzing, and combining resources to develop ideas, make decisions, and design products. Thus, tools that can help knowledge workers track and re-find information is highly desirable.

We believe that provenance is a valuable resource for identifying related documents. Our study had several key findings to support our hypothesis. First, by recording provenance events for 17 knowledge workers over several weeks, we were able to show that about half of a worker's resources are connected with at least one other resource in their repository through provenance. Second, we were able to show that provenance events occur frequently and were memorable to our subjects. Third, our interviews confirmed that provenance can connect related documents, and that graphs produced by such connections give workers a context for reasoning about their work process.

## FUTURE WORK

In addition to an imminent study of provenance and information flow within workgroups, the next step in our work is to examine ways to use provenance information in the user interface. Though some tools are available which use provenance to aid desktop search, we believe this information could more effectively be used as part of file/folder listings in addition to standard thumbnails, detailed listing, etc. This would allow users to search for files through orienteering and navigation—which is preferred over keyword search—while automatically identifying related resources through provenance relationships.

We are also in the process of preparing the system-logging tool for wider use. The goal is to make it easier for researchers interested in all aspects of provenance to more easily gather and use real-world data in their research. Researchers interested in prototyping novel interface elements and ways of visualizing provenance will be able to find a resource to get them started without having to do all the back-end work. At the same time we wish to invite other researchers and developers to help us further develop this logging platform. In addition to addressing the shortcomings already discussed in this paper, a number of new applications need to be instrumented, including open source competitors to Microsoft's applications, increasingly growing in popularity.

There are of course risks associated with releasing this kind of application more widely. It is easy to imagine how this type of a logging application could be misused, and installed on an unsuspecting users' machine. This is something that we would hate to see, but whether there are way to prevent this type of application from seeing this type of use is questionable.

## ACKNOWLEDGMENTS

We thank the many people who helped make this study possible, including the TaskTracer team; Tom Dietterich, John Herlocker, Michael Slater, Jed Irvine, Simone Stumpf, Jill Cao, Ben Porter, Balaji Lakshminarayanan, Xinlong Bao, and Erin Fitzhenry. We also thank our Intel Corporation collaborators, Catherine Spence, and participants for giving so generously of their time and putting up with technical glitches. This work is in part supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. FA8750-07-D-0185/0004 and a gift from the Intel Corporation.

## REFERENCES

1. Barreau, D. and Nardi, B. A. 1995. Finding and reminding: file organization from the desktop. SIGCHI Bull. 27, 3 (Jul. 1995), 39-43
2. Bergman, O., Beyth-Marom, R., Nachmias, R., Gradovitch, N., and Whittaker, S. 2008. Improved search engines and navigation preference in personal information management. ACM Trans. Inf. Syst. 26, 4 (Sep. 2008), 1-24.
3. Blanc-Brude, T. and Scapin, D. L. "What do people recall about their documents?: implications for desktop search tools". In Proceedings of the 12th international

Conference on Intelligent User interfaces IUI '07. ACM, New York, NY, P.102-111, 2007

4. Boardman, R. and Sasse, M. A."Stuff goes into the computer and doesn't come out": a cross-tool study of personal information management. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems CHI '04. ACM, New York, NY, P.583-590, 2004.
5. Bureau of Labor Statistics. U.S. Department of Labor. "Occupational Employment and Wages." Press release. Washington, D.C. 1 May 2009.
6. Chau, D., Myers, B., and Faulring, A. "What to do when search fails: finding information by association". In Proceeding of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems CHI '08. ACM, New York, NY, P.999-1008, 2008
7. Cheney, J., Buneman, P., and Ludäscher, B. 2008. Report on the Principles of Provenance Workshop. SIGMOD Rec. 37, 1 (Mar. 2008), P.62-65, 2008
8. Chirita, S. Costache, W. Nejdl, and R. Paiu. Beagle++: "Semantically enhanced searching and ranking on the desktop". In ESWC 2006, pages P.348—362, 2006
9. Corbin, J. and Strauss, A. 1990. Grounded Theory Research: Procedures, Canons, and Evaluative Criteria. In Qualitative Sociology, Vol. 13, No. 1, 1990.
10. Czerwinski, M. and Eric Horvitz, E. "An investigation of memory for daily computing events". In HCI 2002, pages 230–245, London, England, 2002.
11. Czerwinski, M., E. Horvitz, E., and S. Wilhite, S. A "Diary Study of Task Switching and Interruptions", Proceedings of CHI 2004, ACM Conference on Human Factors in Computing System , 2004.
12. Davis, G. B. 2002. Anytime/anyplace computing and the future of knowledge work. Commun. ACM 45, 12 (Dec. 2002), 67-73.
13. Dragunov, A.N., Dietterich, T.G., Johnsrude, K., McLaughlin, M., Li, L. and Herlocker, J., "TaskTracer: A Desktop Environment to Support Multi-tasking Knowledge Workers", In Proc. IUI'05. ACM Press P.75-82., 2005
14. Drucker, Peter F. 1999. "Knowledge-Worker Productivity: THE BIGGEST CHALLENGE." California Management Review 41, no. 2: 79-94. Business Source Premier, EBSCOhost (accessed May 9, 2009).
15. González,V.M and Mark, G. "Constant, constant, multi-tasking craziness: managing multiple working spheres" In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems" CHI 2004, ACM, New York, NY, P. 113-120, 2004, 2004
16. Gonçalves, D. and Jorge, J. A. "Describing documents: what can users tell us?". In Proceedings of the 9th international Conference on intelligent User interfaces. IUI '04. ACM, New York, NY, P.247-249. 2004
17. Gonçalves, D. and Jorge, J. A. In search of personal information: narrative-based interfaces. In Proceedings of the 13th international Conference on intelligent User interfaces IUI '08. ACM, New York, NY, P.179-188, 2008
18. Jaccard, P. "Etude comparative de la distribution florare dans une portion des Alpes et des Jura", Bulletin del la Société Vaudoise des Sciences Naturelles 37, P.547-579, 1901
19. Kidd, A. 1994. The marks are on the knowledge worker. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Celebrating interdependence (Boston, Massachusetts, United States, April 24 - 28, 1994). B. Adelson, S. Dumais, and J. Olson, Eds. CHI '94. ACM, New York, NY, 186-191.
20. "Provenance." Merriam-Webster Online Dictionary. 2009. Merriam-Webster Online. 28 May 2009 <http://www.merriam-webster.com/dictionary/provenance>
21. Ravasio, P., Schär, S. G., and Krueger, H. In pursuit of desktop evolution: User problems and practices with modern desktop systems. ACM Trans. Comput.-Hum. Interact. 11, 2 (Jun. 2004), P. 156-180, 2004
22. Soules, C. A. and Ganger, G. R. 2005. Connections: using context to enhance file search. SIGOPS Oper. Syst. Rev. 39, 5 (Oct. 2005), 119-132
23. Soules, C. and Ganger, G. Using context to assist in personal file retrieval. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, 2006
24. Stumpf, S., Fitzhenry, E., Dietterich, T.G. (2007) The Use of Provenance in Information Retrieval, Workshop on Principles of Provenance (PROPR), Edinburgh, Scotland, 19-20 November, 2007.
25. Teevan, J., Alvarado, C., Ackerman, M. S., and Karger, D. R. 2004. The perfect search engine is not enough: a study of orienteering behavior in directed search. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Vienna, Austria, April 24 - 29, 2004). CHI '04. ACM, New York, NY, 415-422.
26. Tulving, E. and Thomson, D., "Encoding Specificity and Retrieval Processes in Episodic Memory," Psychological Review, Vol. 80, No. 5, 352-373, 1973