AN ABSTRACT OF THE THESIS OF

Heather Lonsdale for the degree of Master of Science in Computer Science presented on June 8, 2009.

Title: Using Provenance to Aid Document Re-finding.

Abstract approved:

________________________________________
Carlos Jensen

In the field of Human-Computer Interaction, provenance refers to the complete history and genealogy of a document. Provenance can be useful in identifying related resources, such as different versions of the same document or resources used in the creation of a new document. Though methods of provenance collection and applications for provenance have been studied, no studies have documented the frequency of provenance events in typical computer use. We conducted a study of knowledge workers at Intel and used event-logging software to track provenance events as they occurred in the workplace over several weeks. We also interviewed knowledge workers to evaluate the effectiveness of provenance cues for document recall. Our data shows that provenance relationships are quite common, and provenance helps users recall more about their documents and understand the context of their workflows. Through a detailed analysis of the challenges facing knowledge workers, their typical work practices, and the utility of provenance, we argue that provenance can be useful in applications like desktop search.
Using Provenance to Aid Document Re-finding

by
Heather Lonsdale

A THESIS
submitted to
Oregon State University

in partial fulfillment of
the requirements for the
degree of
Master of Science

Presented June 8, 2009
Commencement June 2010
Master of Science thesis of Heather Lonsdale presented on June 8, 2009.

APPROVED:

Major Professor, representing Computer Science

Director of the School of Electrical Engineering and Computer Science

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Heather Lonsdale, Author
ACKNOWLEDGEMENTS

None of this work could have been done without the TaskTracer team. I would especially like to thank Jill Cao, Ben Porter, and Michael Slater for their significant efforts in coordinating and carrying out the study at Intel.

I would also like to thank my parents, Randy and Ellen, and my sister, Dana, for their endless encouragement throughout college. I’m lucky to have such an amazing and supportive family behind me.

To my childhood friends and the friends I’ve made at OSU, thank you all for the good times and exciting adventures. I’d go crazy without you guys!

Perhaps most importantly, I would like to thank my advisor, Carlos, for being a great mentor and friend throughout my years at OSU. I couldn’t be here without your guidance and support.

Research was supported by a grant from Intel Corporation IT Research and Development Council, and end user studies were conducted onsite at Intel Oregon sites.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>2. Challenges faced by knowledge workers</td>
<td>4</td>
</tr>
<tr>
<td>2.1. A day in the life</td>
<td>6</td>
</tr>
<tr>
<td>2.2. Multitasking and interruptions</td>
<td>8</td>
</tr>
<tr>
<td>2.3. Across teams and time zones</td>
<td>9</td>
</tr>
<tr>
<td>2.4. Information overload</td>
<td>11</td>
</tr>
<tr>
<td>2.4.1. Email</td>
<td>12</td>
</tr>
<tr>
<td>2.4.2. Documents</td>
<td>13</td>
</tr>
<tr>
<td>2.4.3. Web</td>
<td>14</td>
</tr>
<tr>
<td>3. Coping strategies</td>
<td>17</td>
</tr>
<tr>
<td>3.1. Work practices: it's all in your head</td>
<td>17</td>
</tr>
<tr>
<td>3.2. Information reuse</td>
<td>18</td>
</tr>
<tr>
<td>3.3. Provenance</td>
<td>19</td>
</tr>
<tr>
<td>3.3.1. Provenance collection</td>
<td>20</td>
</tr>
<tr>
<td>3.3.2. Provenance and memory</td>
<td>21</td>
</tr>
<tr>
<td>3.3.3. Provenance-based search tools</td>
<td>22</td>
</tr>
<tr>
<td>3.3.4. Provenance graphs</td>
<td>24</td>
</tr>
<tr>
<td>4. Experiment</td>
<td>26</td>
</tr>
<tr>
<td>4.1. Goals and motivation</td>
<td>26</td>
</tr>
<tr>
<td>4.2. Study design</td>
<td>27</td>
</tr>
</tbody>
</table>
# TABLE OF CONTENTS (Continued)

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5. Results</td>
<td>30</td>
</tr>
<tr>
<td>5.1. Participant characteristics</td>
<td>30</td>
</tr>
<tr>
<td>5.2. A day in the life: revisited</td>
<td>32</td>
</tr>
<tr>
<td>5.3. Provenance in the workplace</td>
<td>37</td>
</tr>
<tr>
<td>6. Discussion</td>
<td>48</td>
</tr>
<tr>
<td>7. Conclusion</td>
<td>51</td>
</tr>
<tr>
<td>Bibliography</td>
<td>53</td>
</tr>
<tr>
<td>Figure</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
</tr>
<tr>
<td>1. Example of how resources can be related to each other.</td>
<td>2</td>
</tr>
<tr>
<td>2. Average percentage of time spent on activities [Gonzáles, 2004].</td>
<td>6</td>
</tr>
<tr>
<td>3. How knowledge workers use the web [Sellen, 2002].</td>
<td>15</td>
</tr>
<tr>
<td>4. Screenshot of TaskTrail user interface [Stumpf, 2007].</td>
<td>24</td>
</tr>
<tr>
<td>5. Age distribution of the 24 subjects.</td>
<td>30</td>
</tr>
<tr>
<td>6. Distribution of file types attached to and saved from emails.</td>
<td>32</td>
</tr>
<tr>
<td>7. Distribution of used resources by file type.</td>
<td>34</td>
</tr>
<tr>
<td>8. Document focus duration (normalized, natural log scale).</td>
<td>36</td>
</tr>
<tr>
<td>9. Distribution of provenance events by type.</td>
<td>38</td>
</tr>
<tr>
<td>10. Sources and destinations of copy-and-paste by file type.</td>
<td>40</td>
</tr>
<tr>
<td>11. Information flow between Microsoft Office applications.</td>
<td>41</td>
</tr>
<tr>
<td>12. Information flow out of Outlook and IE based on copy/paste.</td>
<td>41</td>
</tr>
<tr>
<td>13. Actual provenance graph for one of our subjects.</td>
<td>42</td>
</tr>
<tr>
<td>14. Common interview elements and percentage of interviewees who mentioned them.</td>
<td>44</td>
</tr>
<tr>
<td>15. Provenance graph that helped trigger a subject’s memory.</td>
<td>47</td>
</tr>
<tr>
<td>16. Mockup of provenance viewer integrated into folder navigation.</td>
<td>50</td>
</tr>
</tbody>
</table>
LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Our division of occupations into knowledge work and others</td>
<td>5</td>
</tr>
<tr>
<td>2. Provenance event counts by application type and relationship</td>
<td>39</td>
</tr>
</tbody>
</table>
1. Introduction

Computer work practices have changed dramatically since the days of the first personal computer. Due to the availability of cheap and abundant storage, our desktops and laptops have become black holes of electronic resources (files, emails, webpages, and contacts). Consequently, finding a specific resource can be a significant challenge. Though several desktop search tools are available, such as Google Desktop or Mac OSX Spotlight, computer users typically prefer manual folder searching over keyword-based search tools [Barreau, 1995; Bergman, 2008; Teevan, 2004]. As an alternative, we suggest tracking relationships between electronic resources and showing these relationships to help users find what they are looking for.

There are several ways that resources can be related. For example, say you are developing a PowerPoint presentation with a group of peers. You might receive a draft of the presentation as an attachment to an email, thus the draft is related to that email through an “attachment save” relationship. You may copy information from a webpage into the presentation, and now the webpage is also related to the presentation by a “copy-paste” relationship. If you save the presentation as a new name, the older version is related to your new version through a “save as” relationship, and so on (see Figure 1). We refer to these relationships, which describe the history of a document’s transformations and related resources, as
Provenance comes from a French word meaning “to originate”, and refers to the history of an object’s origin or derivation [Merriam-Webster Online Dictionary, 2009]. In this paper, we specifically focus on provenance as it pertains to digital resources.

Figure 1. Example of how resources can be related to each other.

Provenance has only recently caught the attention of the research community, and little research has been done on provenance relationships in everyday use. The primary reason for this is that it requires dynamic event monitoring, or observing what the user does on his or her computer and logging provenance events. Current operating systems such as Windows, Mac OSX, and Linux do not track provenance relationships, so researchers have developed their own tools for monitoring and recording provenance. These tools are promising, but
many of them have only been evaluated by their developers or a small select sample. Since event monitoring is dynamic, users also must run the software for a significant period of time in order to build up a database of provenance relationships. Furthermore, none have reported on the frequency of different types of provenance relationships.

To build the case for using provenance to assist in document re-finding, we conducted a longitudinal study at Intel. We received questionnaires from 24 employees about their demographics and work practices and ran provenance event logging software on 17 employees’ work computers for several weeks. We then used the provenance data to interview 15 users individually about their documents and test how well they could recall document attributes. We also observed 9 subjects in their workplace environment to learn more about their work practices and anything that may not be logged by the computer. Through the questionnaire, data logging, interview, and observation, we were able to gather rich data about work practices and provenance.
2. Challenges faced by knowledge workers

In today's workplace, knowledge workers are highly influential to a company's success and growth. 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 possess and their ability to apply knowledge and develop new understanding [Kidd, 1994; Zhou, 2007; Davis, 2002]. While 20th century productivity centered around manual labor, the 21st century largely depends on the success of knowledge workers. Peter Drucker, who originally coined the term “knowledge worker”, argued that “the most valuable asset of a 21st-century institution (whether business or non-business) will be its knowledge workers and their productivity” [Drucker, 1999].

The North American Industry Classification System (NAICS) defines the information sector as “(1) those engaged in producing and distributing information and cultural products; (2) those that provide the means to transmit or distribute these products as well as data or communications; and (3) those that process data” [“2007 NAICS Definition,” 2007]. Despite this somewhat narrow definition, there are over 3 million information sector employees (2.2% of the workforce) in the United States [Bureau of Labor Statistics, 2009].

The more commonly applied definition of knowledge work is much broader than the NAICS definition, and includes those who acquire and apply knowledge in their daily work activities. Most information technology (IT) positions fall into the
Table 1. Our division of occupations into knowledge work and others.

<table>
<thead>
<tr>
<th>Knowledge Work</th>
<th>Non-Knowledge Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Management</td>
<td>Community and social services</td>
</tr>
<tr>
<td>Business and financial operations</td>
<td>Arts, design, entertainment, sports, and media</td>
</tr>
<tr>
<td>Computer and mathematical science</td>
<td>Healthcare support</td>
</tr>
<tr>
<td>Architecture and engineering</td>
<td>Protective service</td>
</tr>
<tr>
<td>Life, physical, and social science</td>
<td>Food preparation and serving related</td>
</tr>
<tr>
<td>Legal</td>
<td>Building and grounds cleaning and maintenance</td>
</tr>
<tr>
<td>Education, training, and library</td>
<td>Personal care and service</td>
</tr>
<tr>
<td>Healthcare practitioner and technical</td>
<td>Sales and related</td>
</tr>
<tr>
<td>Office and administrative support</td>
<td>Farming, fishing, and forestry</td>
</tr>
<tr>
<td></td>
<td>Construction and extraction</td>
</tr>
<tr>
<td></td>
<td>Installation, maintenance, and repair</td>
</tr>
<tr>
<td></td>
<td>Production</td>
</tr>
<tr>
<td></td>
<td>Transportation and material moving</td>
</tr>
</tbody>
</table>

category of knowledge work, including managers, programmers, analysts, and accountants [Davis, 2002]. Other fields such as design, advertising, marketing, and law, are also examples of knowledge work [Kidd, 1994]. Using this definition, we estimate that knowledge workers make up 43.78% of the workforce (see Table 1) [Bureau of Labor Statistics, 2009]. Regardless of which definition one adheres to, or how one segments the workforce into knowledge workers and others, knowledge workers are a very important and growing part of today’s service-oriented economy. They are integral to the way we live and work, and to the advancement of science and technology.
The purpose of this chapter is to describe the complex environment surrounding knowledge work. We begin by describing a day in the life of knowledge workers, including common work practices and how time is spent. This is followed by research on multitasking behavior and interruptions. Next, we describe the prevalence of cross-team and cross-time zone work. Finally, we discuss evidence of information overload experienced by knowledge workers.

### 2.1. A day in the life

Unlike manual workers, knowledge workers are usually the ones responsible for managing their own work. Consequently, work styles can vary greatly between individuals. However, if one thing is certain, it is that a knowledge worker’s productivity, quality of work, and on-time completion depends greatly on their ability to manage their time, attention, and motivation [Davis, 2002].

![Figure 2. Average percentage of time spent on activities [Gonzáles, 2004].](image-url)
Figure 2 shows how knowledge workers typically distribute their time. Deskwork undoubtedly consumes most of a typical knowledge worker’s day, and the average time spent on solitary deskwork is approximately three hours each day [Su, 2008; Gonzáles, 2004]. Knowledge workers also spend over three hours each day (37% of their day) on the computer, which makes up the vast majority of device usage when compared to paper documents, books, calculators, planners, phones, and other devices. Approximately 2 hours of each day is spent in meetings [Su, 2008; Gonzáles, 2004]. One study of 24 knowledge workers estimated that people average 10 meetings each week, of which 5 are scheduled [Bellotti, 2000].

Communication outside of meetings, whether through phone, email, or other mediums, is also a central part of the knowledge worker’s typical day. Most communication occurs in chains rather than single actions [Su, 2008]. In a study of communication chains, people spent an average of 2 hours and 18 minutes in communication acts outside of meetings. Although email is a significant communication medium (see 2.4.1), people also make approximately 10 phone calls each day [Bellotti, 2000].

Although these three major activities each take up a considerable amount of time, they are rarely done independently. Rather, knowledge workers are constantly switching tasks and accepting interruptions, as described in the next section.
2.2. Multitasking and interruptions

At any given time, knowledge workers typically have several simultaneous projects that they must manage. Gonzáles and Mark defined these high-level tasks as "working spheres" [González, 2004]. The primary projects that an individual works on and is held accountable for are referred to as "central" working spheres, while other projects which are not the individual's primary responsibility are considered "peripheral" working spheres. Having more projects has been associated with feelings of poor task coordination with others [Dabbish, 2006].

During a study of 14 information workers at an investment management company, they found that people work on about 6 central and 4 peripheral working spheres on an average day [González, 2004]. Furthermore, information workers spend an average of about 45 minutes each day (or 8% of their day) just managing these working spheres [González, 2004; Mark, 2005]. A diary study of Microsoft employees confirms the need for workers to spend time managing their tasks, with 13% of reported tasks categorized as "task tracking" [Czerwinski, 2004].

A significant challenge for information workers is interruptions and the segmentation of their work. A study of 24 information workers found that people only spend an average of 11 minutes on a “working sphere” before switching to another working sphere or being interrupted [Mark, 2005]. Even within a single task, users tend to switch between mediums very frequently, approximately every three minutes [González, 2004].
Workers experience about 25 interruptions on an average day, half of which are internal (self-initiated) while the other half are external interruptions [González, 2004]. Internal interruptions include using paper documents or the computer, talking to colleagues over cubicle walls, making a phone call, using email, or leaving the cubicle. External interruptions are initiated by others and include new email notifications, in-person visits, and receiving phone calls. In a study at Microsoft, long-term projects were rated significantly more difficult to return to in comparison to short-term tasks [Czerwinski, 2004]. This is likely due to the longer history and greater context that must be managed when returning to a long-term task.

2.3. Across teams and time zones

Knowledge workers must coordinate not only multiple projects, but multiple teams as well. A recent survey at Intel showed that 61% of respondents work with more than 3 teams, and 28% work with more than 5 teams [Chudoba, 2005]. Another survey of 401 current and former MBA students [Mortensen, 2007] indicated that 65% worked on more than one team, confirming the prevalence of multiple team membership. Those belonging to more than one team seemed to devote half of their time to a primary team, with the other half divided amongst the other teams [Mortensen, 2007].

To further complicate team coordination, many people have team members who are not in the same building, or even the same country. The previously mentioned survey of MBA students and graduates reported that 34% of teams were
international, 53% were inter-organizational, and 67% were cross-functional [Mortensen, 2007]. At Intel, approximately 69% of employees collaborate with others in a different time zone, and 13% of employees work in locations different from their managers [Chudoba, 2005]. Although distribution of team members alone had no apparent effect on performance, the variety of work practices used by team members negatively impacted self-perceived performance [Chudoba, 2005].

Membership in multiple teams increases the need for communication and coordination, and requires different methods for collaboration. 70% of the Intel survey respondents reported collaborating with teams without face to face meetings, and 61% work with internet-based conferencing tools [Chudoba, 2005]. In an analysis of communication patterns in the workplace, switching between organizational boundaries during chains of communication was found to be the most predictive factor in job strain [Su, 2008].

Working with remotely located teams can also affect work styles and attitudes. Data has shown that even people who know each other and have worked together before cannot produce the same quality of work when using remote tools [Olson, 2000]. Some studies even suggest that remotely located teams may cooperate less and present themselves more deceptively when compared to collocated teams [Bradner, 2002]. Despite efforts to use tools to support remote teams, work is often reorganized to decrease the collaboration needed with remote team members [Olson, 2000].
2.4. Information overload

It’s no surprise that one of the most significant challenges for knowledge workers is information overload. Most knowledge workers say information overload is intrinsic to their jobs [Mulder, 2006]. When the amount of information at hand is more than one can process, the feeling of information overload can lead one to feel stressed and overwhelmed. Since most of the tasks of knowledge workers are open-ended, it becomes difficult to decide what information is necessary and what is superfluous, resulting in an avalanche of information.

Several factors contribute to feelings of information overload. The most obvious is the vast amount of files, emails, websites, and other resources that one interacts with on a daily basis. Project resources tend to span over different types (for example, documents, emails, and websites), forcing users to manage each type in its own environment [Bergman, 2006]. Managing these resources introduces significant overhead, and few users spend time organizing their resources beyond an occasional “spring-clean” [Boardman, 2004].

Almost by definition, knowledge work involves (and perhaps thrives on) continuous information overload. One study defined the main elements of information overload as (1) task complexity, (2) information ambiguity, (3) environment distractions, and (4) social pressure, such as handling requests from coworkers [Mulder, 2006]. We have already discussed (3) and (4) in the previous sections of this chapter. The rest of this section addresses the other two aspects of
information overload in terms of the volume of resources that knowledge workers must handle.

2.4.1. Email

The advent of email has expanded the ways people can communicate, while also expanding the amount of information people must deal with. When knowledge workers were asked to report incidents of information overload, the majority of reported incidents were related to email [Mulder, 2006]. Some factors contributing to increased feelings of email overload include perceived importance of email for work, number of meetings per week, number of subordinates, overall email volume, and proportion of messages which are “spam” [Dabbish, 2006]. Studies have also shown email being used for personal information management, with most subjects reporting using email to send themselves reminders [Ducheneaut, 2001].

There have been many studies on email volume in the workplace. Some estimate incoming email volume to be as much as 67 messages each day [Bellotti, 2000], though others have reported averages closer to 40 messages [Dabbish, 2006; Ducheneaut, 2001]. Average inbox sizes vary greatly, particularly since some people keep all their messages in the inbox without filing. Regardless of the numbers, knowledge workers receive a considerable amount of email each day, which contributes to information overload.

The standard strategy for email management is through folders. People tend to categorize their folders by role, project, sender, organization, or interest
[Boardman, 2004; Ducheneaut, 2001]. Having a greater number of email folders has been associated with greater email experience [Ducheneaut, 2001], but also with increased feelings of email overload [Dabbish, 2006]. Folder management strategies have been classified as frequent filers, extensive filers, partial filers, and no-filers [Boardman, 2004].

Another strategy for email management involves how frequently incoming email is monitored. Though checking incoming messages as soon as they are received can distract from the current task, it has also been associated with lower feelings of email overload [Dabbish, 2006]. On the other hand, some people choose to restrict email checking to specific times in order to avoid distraction. However, this practice is actually associated with increased feelings of email overload [Dabbish, 2006].

Email is also frequently used as a method for file exchange [Ducheneaut, 2001], with approximately 9 attachments received through email each day [Bellotti, 2000]. This contributes to the problem of document overload, which is discussed next.

2.4.2. Documents

Knowledge workers deal with a substantial number of documents for their research. This includes both sources of information and destinations (reports, results, presentations). Although file system capacity has increased dramatically, studies show that file system fullness has remained relatively constant [Agrawal,
This suggests that the number of files people store on their computers increases steadily with file system capacity. Another study estimated that users added 6 files to their computers each day [Boardman, 2004].

Users organize their files more diligently than email or web bookmarks [Boardman, 2004]. However, classifying documents and folders takes considerable cognitive effort. People have an average of 56.6 folders in their primary document collection (i.e. “My Documents”, UNIX home directory, or a network drive), and the average folder depth is 3.3 [Boardman, 2004]. Folders are commonly named by short-term activities, document class, or long-term activities [Boardman, 2004].

Despite this vast number of folders, users still have an average of 66 “unfiled” documents in a root folder or on the desktop [Boardman, 2004].

Document storage and organization is an ongoing process, and no folder structure is considered permanent [Ravasio, 2004]. However, folder structure tends to change very slowly, and users rely on this semi-steady organization for location-based finding [Boardman, 2004]. Users maintain and archive files at milestones (such as project completion), and archived (non-working) files range in age from 6 months to 8 years [Ravasio, 2004]. Otherwise, extensive file and folder maintenance only occurs at major life stages, like beginning a new job [Boardman, 2004].

**2.4.3. Web**

Of the three “information overload” collections we’ve discussed, web bookmarks tend to be the least organized [Boardman, 2004]. However, knowledge
workers tend to use bookmarks more often than other users in order to keep track of trusted, frequently used information sources [Sellen, 2002].

A study of how knowledge workers use the internet identified three primary categories of web activities: finding, information gathering, and browsing [Sellen, 2002]. “Finding” involved looking for a specific fact, “information gathering” included more general research on a specific topic, and “browsing” occurred when a person went to a site with no specific purpose other than to be informed, updated, or entertained. Transacting, housekeeping, and communicating were other activity categories identified, but each made up 5% or less of the total activities (see Figure 3). Another study identified four modes of information seeking on the web: undirected viewing, conditioned viewing, informal search, and formal search [Wei Choo, 2000].

Figure 3. How knowledge workers use the web [Sellen, 2002]
Knowledge workers play a key role in our society, and the complex nature of their work makes them worthy of further study. The challenges they face through multitasking, multi-teaming, and information overload force them to employ strategies in order to manage their work. The next chapter discusses some of the common coping strategies of knowledge workers.
3. Coping strategies

Knowledge workers are extremely diverse in both their expertise and work styles, and each has a unique way of dealing with the previous mentioned challenges of knowledge work. Rather than trying to cover all of these strategies, this chapter focuses on the general work practices of knowledge workers and the ways they reuse information. We then introduce the idea of provenance as a way to manage and track information reuse.

3.1. Work practices: it's all in your head

After studying knowledge workers, Alison Kidd concluded: “the marks which can make a difference to their organizations are on the knowledge workers not on the pieces of paper” [Kidd, 1994]. This statement remains very true, as it is impossible to include all relevant knowledge and understanding in any written artifact. As a result, knowledge workers have their most valuable asset on them at all times—all in their heads.

Despite being caught in constant information overload, knowledge workers tend to leave that baggage at their desks. Most do not carry much with them other than a couple of notebooks [Bellotti, 2000; Kidd, 1994]. They take a lot of notes, but mostly for short-term use. Only about one-third file notes separately from their notebooks [Bellotti, 2000].

Since knowledge workers cannot predict what will inform them or how they might use a document in the future, they tend to delay filing information until they
are certain they will not need it again [Kidd, 1994]. Whether the clutter collects on
the physical or computer desktop, they also use the physical layout to remember
where they left off or demonstrate progress that has been made [Kidd, 1994].
Knowledge workers also tend to organize their own documents more than
documents authored by others [Boardman, 2004].

3.2. Information reuse

Though much of the activity of knowledge workers happens in their heads,
the documents and reports that they produce are rarely from scratch. Parts of other
documents may be compiled to create a new document, or a template may be used
as a starting point. Levy identified four processes by which new documents come
into being: creation, collection, combination, and customization [Levy, 1993].
Creation is the only process that introduces all-new material; collection involves
gathering existing information from multiple sources, while combination includes
both new and existing material. Customization is when existing material is tailored
for a new purpose or setting.

Situations of information reuse can also be categorized by the characteristics
of the person who is reusing knowledge, and their purpose for doing so. Knowledge
“re-users” fall into four categories: shared work producers, shared work
practitioners, expertise-seeking novices, and secondary knowledge miners [Markus,
2001]. Shared work producers work in teams to produce knowledge for their own
use, and their information reuse is a collaborative effort. On the other hand, shared
work practitioners work independently and reuse knowledge produced by others in similar work roles. Expertise-seeking novices are those who occasionally need information outside of their primary work role, and thus reuse knowledge that they themselves have no experience with. Finally, secondary knowledge miners use knowledge produced by others for purposes much different than their original use; for example, to answer a new question with results of previous research.

No matter how information is reused, the final product of the information reuse process has a “history” of where pieces of information came from and how they got there. We refer to elements of a document’s history as provenance, which we explain next.

3.3. Provenance

Provenance is traditionally defined as “the history of ownership of a valued object” [Merriam-Webster Online Dictionary, 2009]. In terms of document provenance, we consider not only ownership history, but also the series of transformations and sources of information that produce a document. Provenance is critical in understanding the authenticity, integrity, and trustworthiness of a document [Cheney, 2008]. We also believe it can be useful for connecting related resources and improving document re-finding.

The research community has only recently begun addressing provenance in terms of document history. Provenance-Aware Storage Systems (PASS), a research group at Harvard University, has been the most active in this area. The University of
Pennsylvania also hosted a workshop on provenance in 2007, which discussed the technical aspects of provenance collection and possibilities for provenance use in the future [Cheney, 2008]. The biggest hurdle for provenance researchers has been methods for collecting and storing the vast amounts of provenance data that can be produced.

### 3.3.1. Provenance collection

Standard computer operating systems do not currently track provenance. Researchers develop their own software for recording provenance on the computer, either at the operating-system level or the application level. Here we discuss the methods and limitations of both.

The PASS group argues that provenance should be collected at the system level, so that the same entity manages both file storage and provenance data [Seltzer, 2005]. Their implementation consists of a storage system, an automatic provenance collector, and query tools. The collector records workflow events, interprets those events which signify provenance, and store the provenance information with file metadata. Three types of queries can be performed on the provenance storage system: finding a file’s immediate provenance links, tracing the overall ancestry of a file’s provenance history, or querying for specific provenance attributes. One issue with this method is the level of granularity it produces: although it can detect such fine-grained provenance as changes to a single bit of a file, effort is required to translate system-level events into a bigger picture of
provenance transformation [Braun, 2006]. Application-level provenance tracking tends to produce higher-level provenance relationships that make more sense to a typical computer user.

The provenance collector used in our study is a derivation of TaskTracer, a desktop activity management system [Dragunov, 2005]. TaskTracer uses COM (Component Object Model) add-ins to trace events in Microsoft Office applications (Excel, Word, and PowerPoint), Outlook, and Internet Explorer. Although this limits the provenance we gather to Microsoft applications, studies show that these are still the most widely used productivity applications [Bellotti, 2000; Dabbish, 2006]. Though TaskTracer captures a wide variety of computing events, the provenance-specific events it captures include copy/paste, email attachments (both sending and receiving), internet download and upload, and file operations in Windows Explorer. TaskTracer keeps a record for every resource (files, websites, emails, and folders) in a database and records provenance events with the resource ID of both the source and destination.

3.3.2. Provenance and memory

One major reason for studying provenance is that provenance relationships can be an effective memory cue, and thus may help users in desktop search. Research on how well people remember computing events showed that although people can recall about 18 events after 24 hours, they only remember about 4 of them one month later [Czerwinski, 2002]. In order to assist users in remembering,
we need to identify effective memory cues. Psychology literature has shown that information that was available when a memory was encoded helps when trying to remember that memory later [Tulving, 1973]. We believe that provenance relationships help identify the context of a document and can be effective cues for later retrieval.

Provenance relationships also identify related documents, such as documents where information has been copied from, or emails that the document has been attached to. When asked to describe a document, people can easily recall other related documents [Gonçalves, 2004; Blanc-Brude, 2007]. Our research confirms that related documents, as well as showing how they are related, are effective cues for recall.

### 3.3.3. Provenance-based search tools

Researchers have recognized the value of using provenance relationships in search tools, though these provenance-based tools are not widely used. These tools typically fall into two categories, either using provenance as search query input or as a way to enhance search results.

Feldspar is a search tool that allows users to incrementally specify attributes and relationships of the file they are looking for [Chau, 2008]. It allows users to search for a document by specifying related emails, people, folders, files, webpages, 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, save as, and file system operations require constant system monitoring and are not captured by Feldspar.

Another approach to provenance-based input was based on how people described their documents through storytelling [Gonçalves, 2004]. This search tool, called Quill, suggested fill-in-the-blank sentences to prompt users to specify attributes they remembered about the document they were searching for [Gonçalves, 2008]. Quill’s knowledge base, which is used to retrieve search results, monitors system events and updates information about the user’s documents, email attachments, webpages, applications, and agenda. Although Quill employs system monitoring, it still does not capture copy/paste, save as, and 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 [Soules, 2005]. This tool takes a simple approach to provenance relationships by monitoring file open, read, and write events to detect file input and output. A later descendant of Connections took a more provenance-based approach by detecting causality relationships, viewing each process as taking input to produce output [Soules, 2007]. They determined that users perceived results ranked on causality to be better than locality-based ranking [Soules, 2007].
Beagle++ uses email, web, and document metadata to rank search results based on the number of semantic links, which are similar to provenance relationships [Chirita, 2006]. Like previously mentioned search tools, Beagle++ relies on extracted metadata rather than monitoring user actions.

### 3.3.4. Provenance graphs

In a provenance graph, nodes are resources and directed links between nodes represent the provenance relationship. In the TaskTracer system, TaskTrail is a component of TaskTracer that allows users to view and navigate through their provenance graph [Stumpf, 2007]. However, this is the only tool of its kind, and it has not been evaluated or justified outside of the team that developed it.

![Figure 4. Screenshot of TaskTrail user interface [Stumpf, 2007].](image)

Most of the aforementioned search engines rely on maintaining a relational graph for internal use. The main question that arose through this research was: why
keep the provenance graph in the background? Why not show the users the relationships between these documents to explain why they appeared in search results? Why force users to specify provenance attributes that they remember, rather than showing those links in search results to help the user discern between different documents?

The best explanation for these questions is that people do not understand provenance and prefer more traditional search styles and result listings. However, several studies on search habits have shown that people still prefer to manually search through folders rather than using a search tool because the search process allows them to understand the context of found documents [Barreau, 1995; Bergman, 2008; Teevan, 2004]. We believe that provenance graphs are an excellent way to show the context of documents, which leaves one question remaining: do people understand provenance?
4. Experiment

To learn more about knowledge workers and provenance in the workplace, we conducted an ethnographic study at Intel Corporation. Subjects participated in the study for an average of 8 weeks, during which they continued their normal work practices while our data logging software ran silently in the background. All participants completed a background questionnaire, and a subset also participated in an exit interview and observation session.

This chapter describes the experiment we conducted, and the next chapter presents the results of the experiment.

4.1. Goals and motivation

This was an exploratory study, aimed at gathering basic data to examine the feasibility of recording and the potential use of provenance metadata for augmenting everyday interactions with computers. Although studies have been done on provenance collection [Seltzer, 2005; Braun, 2006] and document recall [Boardman, 2004; Barreau, 1995; Bergman, 2008; Gonçalves, 2008; Gonçalves, 2004; Blanc-Brude, 2007; Chau, 2008; Soules, 2006; Shah, 2007], we are unaware of any studies documenting the frequency of provenance relationships and the utility of provenance cues in document recall. Through our questionnaire, interview, data-logging software, and observation, we were able to gather an abundance of data on information workers at Intel.
Specifically, our goals included: (1) conduct a bottom-up analysis and characterization of information use, provenance, and workflow, (2) investigate the effectiveness of provenance relationships as a memory cue, and (3) see if provenance relationships connect related resources and help indicate project boundaries.

4.2. Study design

We recruited 24 information workers at Intel for the study. Each subject completed a survey about his or her work practices and work style, including information reuse, during an on-site meeting. We then installed a custom activity logging system based on TaskTracer [Dragunov, 2005] on their primary computer. The software was instrumented to record provenance and information access events (document access, moving, saving, copying, pasting, document focus and switching, attachments to emails or web pages). Participants had the option of turning the system off periodically to preserve their privacy, and the data was cleaned for personal and sensitive information before analysis.

Partway through the study, we collected each subject’s data to prepare the exit interviews. From the data, we selected two interesting provenance graphs (see 3.3.4) for each participant and asked participants to tell a story about the documents, similar to the procedure used in [Gonçalves, 2004]. For each graph, we began with a “free-recall” phase, where we presented document names one at a time, without their file extensions, starting with documents more central to the graph and working outwards toward peripheral or “leaf” document nodes in the graph. The number of
documents presented to participants ranged from 4 to 13 documents per set (average: 7.5), depending on the structure and complexity of each graph. After this “free-recall” phase, we had a “cued-recall” phase about the same documents where we showed participants the provenance graph and asked specific questions about document features, to see if these cues helped subjects recall more about the documents. These interviews were conducted either in person or over the phone using screen-sharing software to show participants the documents and provenance graphs.

We also observed 9 of the subjects in their workplace in order to document the context of their activities, as well as identify sources and types of information flow that might be missed by our system. Each of these participants were observed for approximately two hours, during which the researcher sat silently in the participant’s workspace and recorded events and the times they took place. This type of observation is similar to that used in other studies of knowledge workers [González, 2004].

We used an open coding method to analyze the interview and observation logs. This methodology helps us identify patterns and common themes in the logs and is discussed in greater detail in [Corbin, 1990]. Two researchers coded the logs individually, and a third independent 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 total number of codes recorded).
At the end of the study, we collected the data recorded by our logging system and uninstalled the software from their computer.
5. Results

The background questionnaires, software logs, observations, and interviews provided us with a myriad of information to help us understand provenance in the workplace. This chapter describes the quantitative and qualitative results of our study.

5.1. Participant characteristics

We recruited 24 participants from Intel, 10 female (41.7%) and 14 male (58.3%). One subject was over 60 years of age, and the rest were between 30 and 59 years old. Our sample included 12 managers (50%), 4 system analysts (16.7%), 2 system engineers, 2 administrative assistants, and one each of the following: enterprise architect, software architect, human factors engineer, and senior administrative associate.

Although 24 participants began the study, three participants had older hardware/software configurations which were not compatible with our software, one chose to leave the study due to other software issues, one began their sabbatical during the study, and another was excluded from analysis due to database
problems. We also chose to exclude one other participant who received a new computer partway through the study, leaving fragmented and unreliable data. This left us with 17 complete datasets. The excluded subjects (3 women, 4 men) did not introduce significant bias to the final data (41.2% female, 58.8% male).

Subjects with complete datasets ran our activity logging software for an average of 43 workdays (minimum: 21; maximum: 63; standard deviation: 13). We collected data for 126,620 unique resources, or about 7,448 resources per subject (minimum: 3,211; maximum: 17,570; standard deviation: 3,326). This includes Excel, PowerPoint, Word, and text files, emails, websites, contacts, and PDFs.

Though we recorded events for nearly every type of resource, Microsoft Excel, PowerPoint, and Word were the most well-instrumented applications. For these three applications alone, we recorded over 44,000 events across the 17 datasets.

We also tracked email attachments, though these required adding new buttons in Outlook for adding and saving attachments in order for our software to capture the events. We explained the new buttons to the subjects at the beginning of the study through a brief tutorial, but it is likely that the numbers of email attachments are a lower boundary as we could not capture other methods of adding and saving attachments. Nonetheless, we recorded 141 instances (8.3 per user) of users attaching a file to their email and 155 instances (9.1 per user) of users saving email attachments. Attachments added to emails were primarily Excel (41.8%), Word
(30.5%), and PowerPoint files (23.4%). Attachments saved from emails were also primarily Microsoft Office documents, though these were mostly Excel (43.9%) and PowerPoint (42.6%) and fewer Word documents (11%).

![Image](image.png)

**Figure 6.** Distribution of file types attached to and saved from emails.

We captured 142 web downloads and 117 web uploads from Internet Explorer, largely due to heavy SharePoint usage. PowerPoint was the most commonly downloaded and uploaded file type, followed by Excel and Word.

### 5.2. A day in the life: revisited

When asked what they do for their job on a daily basis, all but one participant (out of the 24 who answered our background questionnaire) indicated that meetings were a major part of their work, and meetings were the first thing mentioned for more than half of the participants (15 subjects). Of those who estimated the percentage of their workday spent in meetings (19 subjects), half reported spending
10-30% of their day in meetings, and the other half reported spending 50% or more of their time in meetings. Many subjects noted that they had both face-to-face and virtual meetings. Six participants (25%) also reported preparing presentations as a significant part of their workday.

Email was the second-most mentioned work aspect and was referenced by 15 of the participants. Of those who estimated the percentage of their workday spent on email (13 subjects), most of them (11 subjects, 84.6%) stated that email required 10-30% of their workday. The other two subjects, both in administrative roles, reported much higher email use at 50% and 75% of their work. One of these administrative assistants worked with about 139 emails each day according to our software, which was much higher than the 73.7 emails-per-day average of the overall sample. This includes any email, whether newly received, previously received, or sent. Compared to the incoming email estimates discussed in 2.4.1, the overall average of 73.1 emails per day suggests that at least 6 of these are re-read or sent emails.

Over the course of our study, our subjects worked with an average of 9.5 Excel, PowerPoint, Word, PDF, and text files each day. As shown in Figure 7, most of these were Word documents, and text and PDF files were underrepresented. However, for text and PDF files, our software only tracks Adobe Reader and Microsoft Notepad, so files opened in other programs (such as Notepad++ or a web
browser’s PDF viewer) are not accounted for. Subjects also visited an average of 86 webpages per day, which includes different pages visited within the same website.

![Percentage of Files by Type](image)

**Figure 7.** Distribution of used resources by file type.

We observed 9 of our subjects in their workspace and took notes on their activities throughout a one to three hour time period (the average observation period was 1 hour 48 minutes). The observations took place either before lunch (5 participants) or in the afternoon (4 participants). Subjects were observed in their cubicle for most of the time, except for one participant who was observed during a face-to-face meeting in a conference room. Another participant met with a colleague in the cafeteria for 30 minutes, but was otherwise observed in her office. One observation was interrupted by a 15-minute fire drill, which was excluded from the analysis.

All of our participants experienced at least one interruption during the observation period. Interruptions included new email notifications (13 instances over 3 subjects), phone interruptions (12 instances, 3 subjects), MS Outlook meeting
notifications (6 instances, 3 subjects), in-person office visits (4 instances, 3 subjects), and waiting for the computer (11 instances, 5 subjects). One subject used a secondary desktop computer to continue being productive whenever his laptop was rebooting, operating slowly, or experiencing other problems. However, this became problematic when he opened a form on a website on the desktop and needed a spreadsheet from his laptop to fill in the form. He continued switching between using the laptop and desktop over 40 times to fill out the form while referring to files on his laptop. After about 45 minutes, he finally opened the website on his laptop, only to realize he couldn’t get to the same webpage he had on his desktop.

Another subject used Microsoft OneNote as his information repository. During the observation, he had trouble remembering his passwords, and used Microsoft OneNote as a central password storage location. He also copied data from an internal Intel application to a new note in OneNote.

Microsoft Communicator (an internal instant messaging service) was used heavily by 4 of the observed subjects, and one other subject used it once to get a phone number. Communicator was often used for quick tasks such as finding contact information, checking if a person is available, or letting others know what he or she is currently working on. Others used Communicator while troubleshooting a problem, often with LiveMeeting running to share their screen. Communicator was also used during meetings to communicate without interrupting others.
We also observed four subjects using LiveMeeting and telephone headsets to work with others on a project. This allowed the subject to show others their screen while explaining concepts. Two of these “virtual meetings” were observed in their entirety: one lasted for 51 minutes, and the other lasted for 2 hours and 29 minutes. The other two meetings were longer than 25 minutes and continued beyond the observation period.

Figure 8. Document focus duration (normalized, natural log scale).

Data from our logging system confirms the fragmentation of work found in previous studies (see 2.2). Figure 8 shows that focus times of less than one minute were much more common than longer periods of focus. Participants switched application windows very frequently, and Excel and PowerPoint documents were kept in focus for an average of less than one minute. Subjects tended to keep Word
in focus for much longer, 3.77 minutes on average. We also noted frequent window switching during our observations, especially for different programs related to the same task.

Although a major reason for observing participants was to find sources of provenance that are not captured by TaskTracer, there was little evidence of undocumented provenance. The only missing source of provenance that we observed was transcribing, i.e. recording data from another source by hand, without specifically using copy and paste. This includes the perviously mentioned participant who entered data from the spreadsheet on his laptop into a web form on his desktop. Similarly, other participants transcribed when the source of data was not on their computer (such as information from a phone conversation or physical notes). Subjects also transcribed when the information they needed to copy was in a different format, such as a timestamp that needed to be entered in a spreadsheet with date and time in different columns. Overall, 5 of the 9 observed subjects transcribed at some point during the observation. However, these types of provenance are nearly impossible to detect through software.

5.3. Provenance in the workplace

Our software logged 4,809 provenance events over 17 subjects, illustrated in Figure 9 (not including provenance events within the same resource). Copy and paste was the most common provenance event, occurring an average of 4 times each day. Saving a file to another name (either through file rename or “Save as...“) was
the second most common provenance event, occurring 1.4 times each day. Overall, our subjects performed an average of 6.7 provenance actions per day, and the majority (53.7%) of the Excel, PowerPoint, Word, PDF, and text files that our subjects interacted with were related to at least one other file through provenance.

Table 2 shows the number of occurrences for each provenance event, organized by the source (“FROM”) and destination (“TO”) of events, where applicable.
<table>
<thead>
<tr>
<th></th>
<th>Web Download</th>
<th>Web Upload</th>
<th>Attachment Save</th>
<th>Attachment Add</th>
<th>Copy Paste</th>
<th>File Rename</th>
<th>Save As</th>
<th>Move File</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>FROM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PDF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>11</td>
<td>4</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>Excel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>59</td>
<td>739</td>
<td>81</td>
<td>250</td>
<td>53</td>
</tr>
<tr>
<td>IExplorer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>851</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outlook</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>608</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PowerPoint</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>33</td>
<td>274</td>
<td>48</td>
<td>216</td>
<td>46</td>
</tr>
<tr>
<td>Text</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>34</td>
<td>43</td>
<td>8</td>
<td>25</td>
</tr>
<tr>
<td>Word</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>43</td>
<td>510</td>
<td>82</td>
<td>231</td>
<td>119</td>
</tr>
<tr>
<td>TO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PDF</td>
<td>22</td>
<td>8</td>
<td>4</td>
<td></td>
<td></td>
<td>9</td>
<td>4</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>Excel</td>
<td>30</td>
<td>26</td>
<td>68</td>
<td></td>
<td></td>
<td>983</td>
<td>81</td>
<td>250</td>
<td>53</td>
</tr>
<tr>
<td>PowerPoint</td>
<td>47</td>
<td>66</td>
<td>66</td>
<td></td>
<td></td>
<td>451</td>
<td>51</td>
<td>216</td>
<td>46</td>
</tr>
<tr>
<td>Text</td>
<td>16</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
<td>33</td>
<td>43</td>
<td>8</td>
<td>25</td>
</tr>
<tr>
<td>Word</td>
<td>27</td>
<td>15</td>
<td>17</td>
<td></td>
<td></td>
<td>1552</td>
<td>79</td>
<td>231</td>
<td>119</td>
</tr>
<tr>
<td>Total</td>
<td>142</td>
<td>117</td>
<td>155</td>
<td>141</td>
<td></td>
<td>3028*</td>
<td>258*</td>
<td>705*</td>
<td>263*</td>
</tr>
</tbody>
</table>

*Some totals do not equal the sum of the column because the “FROM” and “TO” rows can overlap for these events.

Table 2. Provenance event counts by application type and relationship.
We also tracked applications that were copied from and pasted into, shown in Figure 10. Though we could detect copy events from websites, emails, Notepad, and Microsoft Office applications, only Notepad and Microsoft Office were instrumented to detect pastes. The internet was the most common source for copy-and-paste, followed by Excel and emails. About half of the recorded copy-and-paste events ended up in Word and one-third were into Excel.

Copy and paste events help demonstrate the flow of information between Microsoft Office applications (Excel, PowerPoint, and Word). Figure 11 expresses these relationships as percentages of overall information flow (for example, copies from PowerPoint to Word made up 13.71% of the overall copy-paste events between MS Office applications). Excel was often a source of information for the other 2 applications, with more than half of copy-paste events coming out of Excel. Subjects also recalled copying from “Excel reports” in many of the interviews. Information
As mentioned previously, we could only track websites and emails as the sources of copying, not as the destinations of pastes. For this reason, information flowing from websites and emails is illustrated separately in Figure 12. The vast majority of data copied from these two applications ended up in Word documents. Keep in mind that the arrows only represent copy-and-paste events, not email attachments or web downloads/uploads.

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 2 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 4 significant resources per workday.

Figure 13 shows one of the graphs we produced by analyzing provenance data. This was also one of the provenance graphs used in the subject’s interview. During the interview, the subject explained that he made a master spreadsheet (circled in Figure 13) to track individual user tests. Each test had a test script saved as an Excel spreadsheet, which then became associated with the master spreadsheet through copy-and-paste events.

Figure 13. Actual provenance graph for one of our subjects.
Interestingly, the subject indicated that he received the test script spreadsheets as email attachments, though our data showed that many of them were created by saving an existing spreadsheet to a different name. This suggests that he may have received some test requests through email and then created new spreadsheets from existing ones based on the requests. Even though the provenance he remembered did not match what our software recorded, other provenance relationships caused these files to still be included in the graph.

However, provenance is not foolproof. Occasionally, provenance connected resources that the subject did not consider to be related. This could be a result of user error (for example, accidentally copying and pasting from the wrong document) or problems with our software. Either way, this was a relatively rare occurrence: only 8 of the 224 files (3.6%) that we interviewed our subjects about were identified as unrelated to the others. On the other hand, every subject mentioned related documents during the interview, most of which were included in the set because they were connected through provenance.

Other relationships were also mentioned in many of the interviews. The most frequently-mentioned aspect of a document was a person who was related to the document, and one-third of the interviewees mentioned remotely-located people associated with a document. 73% of those interviewed mentioned a document being related to a bigger project. Related meetings and presentations were also mentioned by 70% of interviewees.
Figure 14. Common interview elements and percentage of interviewees who mentioned them.
Given only the filename, users could recall the file type correctly 84% of the time, and guessed an incorrect file type the other 16%. Some subjects needed to ask the interviewer for the file type in order to remember anything about the document. Many of the participants admitted that they weren’t very diligent with file naming, and one explained that he relied heavily on the “recent documents” list in programs to find a file.

In general, the types of provenance our subjects recalled in the interviews were those which occurred frequently according to our software. Copy-and-paste, which accounted for 63% of the provenance events we tracked, was mentioned in 87% of the interviews. 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 their interview. Though the remainder of provenance events were fairly evenly distributed, email attachments probably occurred much more often than our data suggests, since we could not track Outlook’s standard attachment buttons (explained at the end of section 5.1). 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 mentions 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, which made up 5.7% of the total specified locations). The majority of document locations (71.4%) were repositories, primarily SharePoint. This resulted in many web-based upload and download provenance events.

The provenance graph was an effective memory cue, helping users recall more information about their documents throughout the interviews. Our subjects also understood what their graphs meant, such as one participant who said, “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 (Figure 15):

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.
Figure 15. Provenance graph that helped trigger a subject’s memory.
6. Discussion

Our study data shows that provenance occurred regularly for our 17 knowledge workers, connected related documents, and could be an effective memory cue. This chapter discusses the implications of these findings and possibilities for future integration of provenance information in the desktop.

This study is one of 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 that are being developed have potential to make an impact in 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.

Our findings also support the anecdotal justifications given for TaskTrail, TaskTracer’s provenance graph visualization tool [Stumpf, 2007]. Though previous evaluations of TaskTrail were from students and faculty in the TaskTracer team, our study shows the potential utility of the tool for knowledge workers. 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 recognize a document in the set or did not see how it was related found it helpful to see the overall provenance graph after seeing each document in the set. 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. As discussed in 2.2, knowledge workers have about 10 different projects each day and spend a significant time just 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 [Barreau, 1995; Bergman, 2008; Teevan, 2004]. For this reason, I 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 the search tools with provenance-based input methods discussed in 3.3.3. Seeing a file in its provenance context helps users distinguish between similarly-named files resulting from provenance relationships such as copying, renaming, or sending and saving email attachments.

Figure 16 is a simple mockup to demonstrate how a provenance graph viewer could be integrated into standard folder navigation. The preview pane in the Mac OSX Finder is replaced by a “relationships” pane, where the selected document is highlighted and shown with provenance-related documents and links. This could be
implemented as another view in the folder browser (in addition to the standard views of thumbnails, list, etc.).

Figure 16. Mockup of provenance viewer integrated into folder navigation.
7. Conclusion

Knowledge workers are a critical component of our society 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 their resources are highly beneficial.

We believe that provenance is a valuable technique 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 have at least one provenance link, and about 6.7 provenance events occur each day. Second, we were able to identify provenance events which occurred more often and were recalled by our interviewees. Third, our interviews confirmed that provenance can connect related documents, and the graphs produced by such connections give workers a context for their workflow. This supplements other recent research that uses machine learning techniques to identify common workflows through provenance graphs [Shen, 2009].

The next step in utilizing provenance is to integrate it into the operating system, especially in folder management (i.e. Windows Explorer or Mac OSX Finder).
Though some tools are available which use provenance to aid desktop search, only TaskTrail shows the user the provenance graph, and we argue that it would be more effective if it were implemented as one of the file/folder listing views in addition to the 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 hope our findings support the work currently being done in provenance, as well as provide guidance for future research. Once provenance is integrated in the operating system, other uses for these relationships will emerge and enhance the desktop experience.
Bibliography


Ducheneaut, N. and Bellotti, V. 2001. E-mail as habitat: an exploration of embedded personal information management. interactions 8, 5 (Sep. 2001), 30-38.


