CHAPTER 4

Sanitization and Visibility 2: Applications

This chapter considers sanitization of information collected while browsing the web and in complex document files. As we saw in Chapter 3, hidden information has resulted in the compromise of security and privacy. We shall also see that the patterns developed at the end of Chapter 3 can be applied to web browsers and document files with similar results.

4.1 Case Study: Sanitizing Web Browser History

It is widely recognized that information retained in web browsers can compromise security and privacy. In part, this is because web browsers record significant information about web pages that have been visited:

- A notation of the page's URL and the time it was visited is kept in the browser's **history**.
- A copy of the page that was downloaded is frequently kept in the browser's **cache**.
- Many web pages download cookies which are stored in the browser's **cookie jar**.

The fact that browsing history is kept in multiple locations is an accident of web browser development. The NCSA Mosaic web browser released in 1994 did not include a persistent history or cache. The Netscape browser introduced the cache to improve browsing performance. However, since the HTTP 0.9 protocol did not have a way to probe the modification time of web objects, the browser's cache could easily become inconsistent. Netscape 1.1 therefore gave the user explicit control over the cache: a preference panel allowed documents to be "verified" once per session, every time the document was downloaded, or "never"—that is, once a document was downloaded, it would not be downloaded again (Figure 4-2). Netscape 1.1 furthermore exposed two caches to the user: a memory cache, with a default of 600 kilobytes in size, and a disk cache, with a default of 5 megabytes. Both could be manually cleared with a button on the Preferences panel.

Netscape 1.1 also came equipped with a rudimentary browser history, as shown in Figure 4-3. But this history was kept in memory and lost whenever the browser was closed. The only way to preserve a history entry was by manually clicking the "Create Bookmark" button.

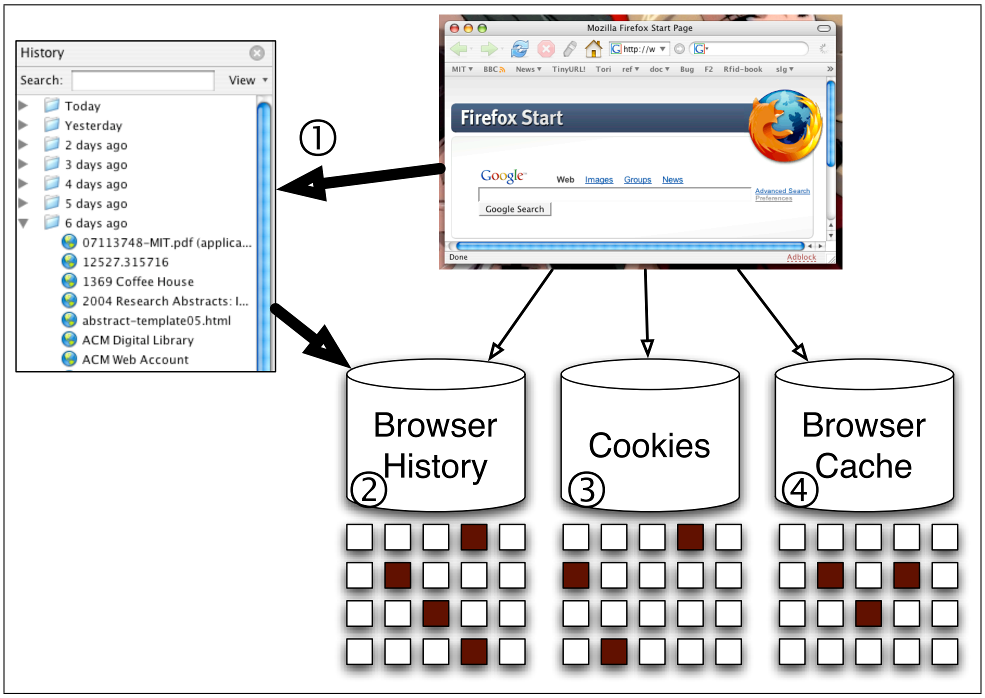


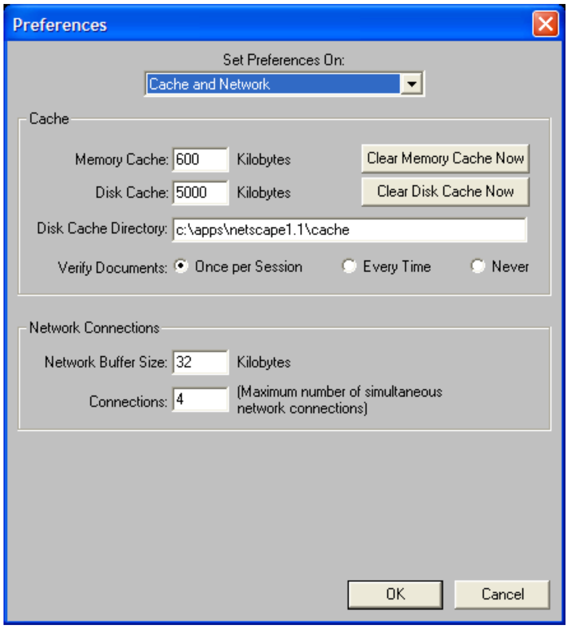
Figure 4-1: The fact that information has been downloaded from a specific web page can be recorded in three places on a modern web browser: in the browser's history, in its cache, and in its cookie files. Even if the files are deleted later, the information may still reside on the computer in recoverable files, illustrated above as shaded boxes.

Modern web browsers employ caches that are considerably larger than the 5 megabytes and keep a persistent history that can go back weeks or longer. There are many reports that this history information has been used to compromise individual's privacy.

Web browsers retain a substantial amount of personal information during the course of normal operation. Information left behind in browsers has also proven to be useful in law enforcement. For example, at the November 2004 murder trial of Michelle Theer, prosecutors introduced forensic evidence including web pages with personal ads that Theer had written in 1999 and web-mail written in responses to those ads, all recovered from web browser files on Theer's computer.[Woo04] Many of the files had been deleted but not yet overwritten. Theer was found guilty on December 3, 2004, of murder and conspiracy and sentenced to life in prison.[WRA04]

Web browsers are in effect data custodians for a significant amount of personal information. Sometimes users are made aware that this personal information is being collected, either through education or through the browser's interface, but it is suspected that many users are not aware of the complete extent of the data collection.

Because web browsers are frequently used on computers shared by more than one person, it is important for browsers to provide users with the ability to remove this personal information when they wish. The American Library Association has adopted a policy that calls for browser history, caches, and cookies to be removed from public access computers in libraries at the end of each day.[Ame05] Even in the case of computers that are not shared, users may still wish to "erase their



Set Preferences On: Cache and Network

Memory Cache: 600 Kilobytes

Disk Cache: 5000 Kilobytes

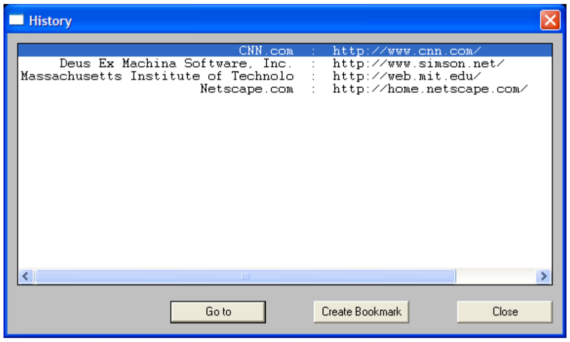
Disk Cache Directory: c:\apps\netscape1.1\cache

Verify Documents: ● Once per Session ○ Every Time ○ Never

Network Buffer Size: 32 Kilobytes

Connections: 4 (Maximum number of simultaneous network connections)

Figure 4-2: Netscape 1.1 “Cache and Network” preference panel gave the user rudimentary control over the browser’s cache.



```
                             CNN.com  :  http://www.cnn.com/
     Deus Ex Machina Software, Inc.  :  http://www.simson.net/
Massachusetts Institute of Technolo  :  http://web.mit.edu/
                       Netscape.com  :  http://home.netscape.com/
```

Figure 4-3: Netscape 1.1’s history was automatically purged after each browsing session. The only way to make a history element persistent was by clicking the “Create Bookmark” button.

tracks” under certain circumstances to guard against the possibility that their computer may be analyzed at a later time by another party.

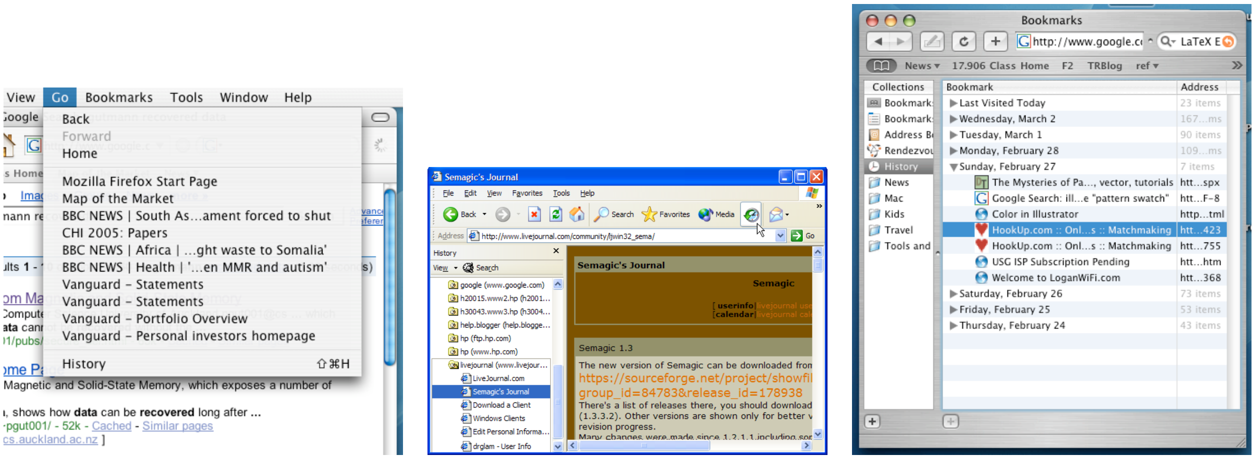
This section considers the alignment of usability and security in three web browsers: Internet Explorer 6.0 (PC), Apple Safari 1.0 (Mac), and Mozilla Firefox 1.0. All of these browsers provide users with various tools for removing information collected during the course of a web browsing session. But these three web browsers take very different approaches. Explorer makes it difficult to remove this information; Safari makes it easy; and Firefox is somewhere in the middle.

What’s more, all of these browsers have a common failure: even when they give the user the ability to delete data, they do not actually remove the data from the computer, because they do not implement COMPLETE DELETE. As the Theer case demonstrates, information can be recovered even if it is not visible from the browser interface.

4.1.1 Web site history

In order for the user to know that information needs to be sanitized, it is first necessary that the user know that the information has been captured. As shown in Figure 4-4, web history information can appear in two different locations in today’s browser. All browsers have the ability to show history in a panel. The Safari and Firefox browsers also have the ability to display history information directly from a menu: in Safari this menu is named “History” while in Firefox the menu is confusingly named “Go.”

Browser designers have adopted two strategies for allowing the user to clear the browser’s his-



The new version of Semagic can be downloaded fron https://sourceforge.net/project/showfil group_id=84783&release_id=178938 There's a list of releases there, you should download (1.3.3.2). Other versions are shown only for better v revision progress. Many changes were made since 1.2.1.1 including sor

Bookmark	Address
▶ Last Visited Today	23 items
▶ Wednesday, March 2	167…ms
▶ Tuesday, March 1	90 items
▶ Monday, February 28	109…ms
▼ Sunday, February 27	7 items
The Mysteries of Pa…, vector, tutorials	htt…spx
Google Search: ill…e "pattern swatch"	htt…F-8
Color in Illustrator	http…tml
HookUp.com :: Onl…s :: Matchmaking	htt…423
HookUp.com :: Onl…s :: Matchmaking	htt…755
USG ISP Subscription Pending	htt…htm
Welcome to LoganWiFi.com	htt…368
▶ Saturday, February 26	73 items
▶ Friday, February 25	53 items
▶ Thursday, February 24	43 items

Firefox history menu IE history pane Safari bookmarks panel

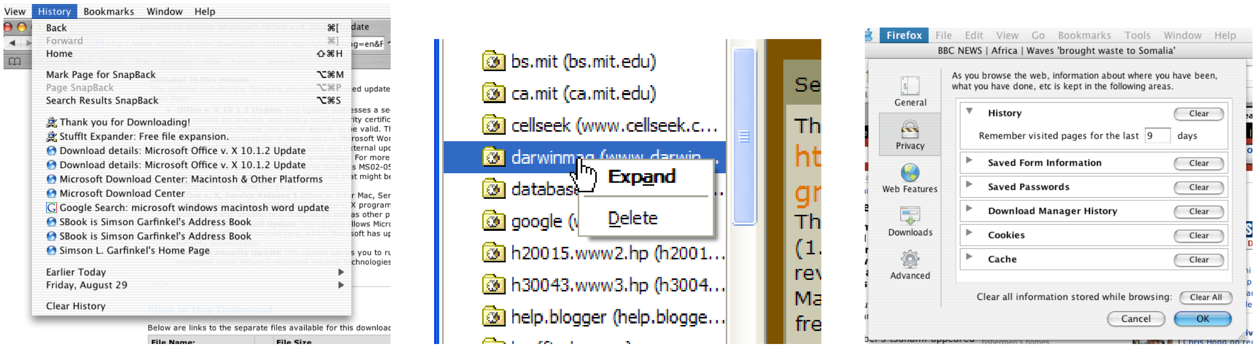
Figure 4-4: History information—the list of web sites that the browser has visited—can appear in two locations of the typical browser interface. The list of web sites can appear directly in the browser's menu, as it does in the Firefox browser (left), or in the IE panel (center). In Safari, the history panel appears as a collection inside the bookmarks panel (right). Display of history is potentially a privacy issue because it can reveal private information about the browser user to other people who have access to the user's computer. In this case, for example, the Safari browser history reveals that the user visited the HookUp.com matchmaking web site.

tory. All of the browsers maintain a list of web pages recently visited that is used to implement the browser "History" feature. Each of the browsers further has a button that can erase this list (Figure 4-6). The browsers also allow individual history items to be eliminated by control-clicking or right-clicking on the specific history item and selecting the "delete" context-menu (Figure 4-5).

Safari's control for clearing the browser history is very easy to find: a menu item clearly labeled "Clear History" is located at the bottom of the "History" menu, as shown on the left in Figure 4-5. Safari presents a control for removing this information where that the information is displayed, an application of the EXPLICIT ITEM DELETE design pattern.

Clearing Explorer's browser history is a multi-step process. First the user must click on the browser's "Tools" menu and select the "Internet Options" menu item. If the Internet Options have been previously displayed and the "General" tab is not selected, it must be selected. Next, the user must click on the "Clear History" button. Finally, the user must confirm the question, "Are you sure you want Windows to delete your history of visited Web sites?" This process is shown visually in Figure 4-6.

Explorer's interface has some significant usability hurdles for an untrained user: the user must know in advance that the "Clear History" button is located on the "Internet Options" panel. The user must realize that having "Windows ... delete your history of visited Web sites?" is the same as clearing Explorer's history menu. Probably the most significant usability problem is that there is no indication on Explorer's History Panel that there is any way to remove this personal information at all! Explorer does *not* follow the EXPLICIT ITEM DELETE pattern. Adding a "Clear History" button to this panel would make the functionality clear.


View
History
Bookmarks
Window
Help
Back
Forward
Home
Mark Page for SnapBack
Page SnapBack
Search Results SnapBack
Thank you for Downloading!
StuffIt Expander: Free file expansion.
Download details: Microsoft Office v. X 10.1.2 Update
Download details: Microsoft Office v. X 10.1.2 Update
Microsoft Download Center: Macintosh & Other Platforms
Microsoft Download Center
Google Search: microsoft windows macintosh word update
SBook is Simson Garfinkel's Address Book
SBook is Simson Garfinkel's Address Book
Simson L. Garfinkel's Home Page
Earlier Today
Friday, August 29
Clear History
bs.mit (bs.mit.edu)
ca.mit (ca.mit.edu)
cellseek (www.cellseek.c...
darwinmag (www.darwin...
database...
google (...
h20015.www2.hp (h2001...
h30043.www3.hp (h3004...
help.blogger (help.blogge...
Expand
Delete
Firefox
File
Edit
View
Go
Bookmarks
Tools
Window
Help
BBC NEWS | Africa | Waves 'brought waste to Somalia'
General
Privacy
Web Features
Downloads
Advanced
History
Saved Form Information
Saved Passwords
Download Manager History
Cookies
Cache
Clear
Clear All
Cancel
OK

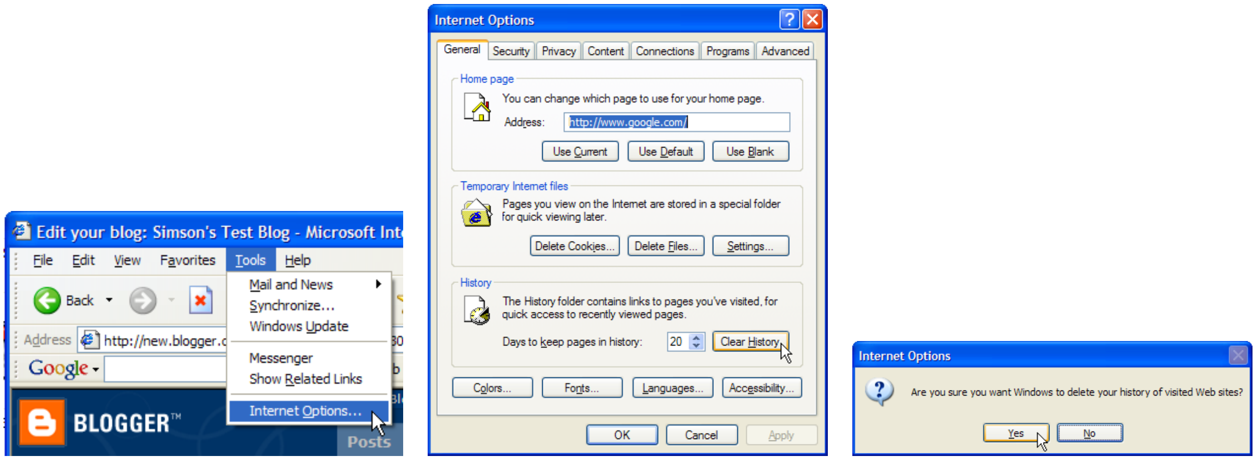
As you browse the web, information about where you have been, what you have done, etc is kept in the following areas.

Remember visited pages for the last 9 days

Clear all information stored while browsing:

Safari Internet Explorer Firefox

Figure 4-5: Different strategies for clearing history information. Safari (left) features a "Clear History" command directly where the history information is displayed. Both Internet Explorer and Firefox allow individual entries in the history panel to be deleted by control-clicking on the history entry (a feature that may not be obvious to many users) Firefox (right) and Internet Explorer also feature a button on the program's preference panel to clear the browser's history—an odd place to put the control, considering that clearing history is not a "preference" that is set.



You can change which page to use for your home page.

Address: http://www.google.com/

Pages you view on the Internet are stored in a special folder for quick viewing later.

The History folder contains links to pages you've visited, for quick access to recently viewed pages.

Days to keep pages in history: 20

Are you sure you want Windows to delete your history of visited Web sites?

IE's Internet Options menu IE's Internet Options panel Clearing IE's History

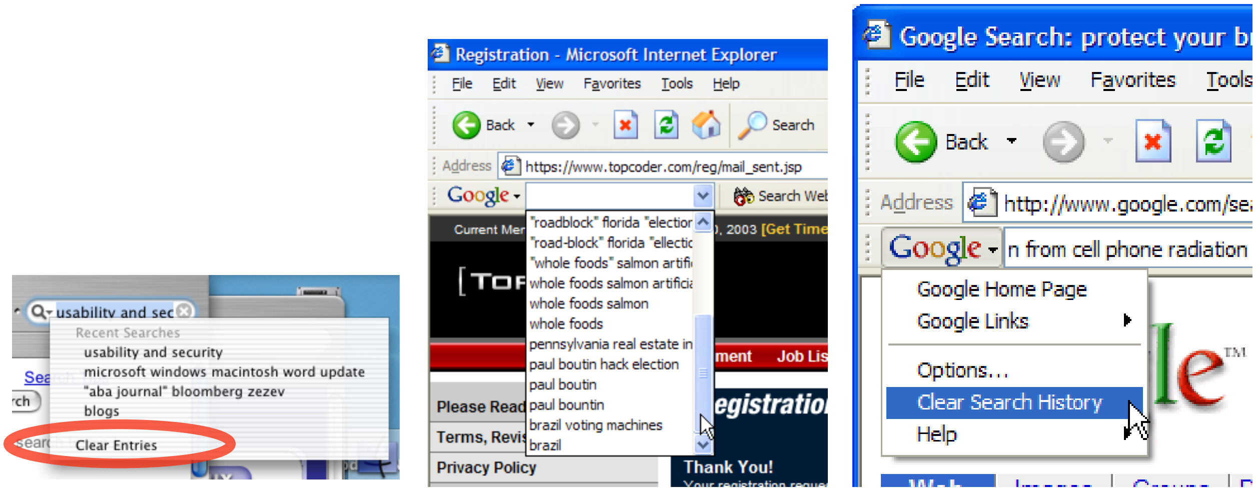
Figure 4-6: Internet Explorer's "Clear History' button is confusingly accessed from the browser's "Internet Options" menu. Selecting the menu option (left) causes the modal "Internet Options" panel (center) to be displayed. Selecting the "Clear History" button causes a modal "Internet Options" alert panel to appear. Clicking "Yes" (right) causes the history files to be unlinked from the Windows file system. The files are not overwritten. Neither the cache files nor the cookies associated with the history pages are altered in any way.

4.1.2 Search history

All three browsers reviewed in this section have the ability to execute a search on the popular Google search engine when the user types a search term into a specially designated field and hits "Enter" or "Return."[1] The Google toolbars remember previous searches so that they can be executed again. These remembered searches are another area where personal information can be compromised.

As with the remembered web history, Safari gives the user a straightforward way to clear the search

[1] Both Safari and Firefox provide this functionality natively, while Internet Explorer requires that the Google Toolbar be separately installed.


usability and sec
Recent Searches
usability and security
microsoft windows macintosh word update
"aba journal" bloomberg zezev
blogs
Clear Entries
Registration - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Address https://www.topcoder.com/reg/mail_sent.jsp
Google
Search Web
"roadblock" florida "election
"road-block" florida "electio
"whole foods" salmon artifi
whole foods salmon artificia
whole foods salmon
whole foods
pennsylvania real estate in
paul boutin hack election
paul boutin
paul bountin
brazil voting machines
brazil
Please Read
Terms, Revis
Privacy Policy
Thank You!
Google Search: protect your b
File Edit View Favorites Tools
Back
Address http://www.google.com/se
Google
n from cell phone radiation
Google Home Page
Google Links
Options...
Clear Search History
Help

Safari Google Toolbar Google Toolbar

Figure 4-7: The list of previously searched terms is another way that history information can be revealed. As with browser history, Safari (left) provides the ability to clear this historical record where the information is displayed. In order to clear the search history of the Google Toolbar (center), the user must select the "Clear Search History" command from the Toolbar's somewhat hidden menu.

history: at the bottom of the list of remembered searches is a menu option that reads "Clear Entries" (Figure 4-7, left).

The Google Toolbar also allows the user to clear the search history, but the approach is more roundabout. Although there is a "Clear Search History" menu command, that command is located under the "Google" menu, rather than at the bottom of the search history (Figure 4-7). Thus, the Google Toolbar does not implement the complete EXPLICIT ITEM DELETE pattern: the ability to delete information is provided, but not where the information is displayed. Once again, adding the ability to delete the information where it is displayed would improve usability by both informing the user that such deletion is possible and giving the user the ability to perform it.

4.1.3 The browser cache: a hidden history

In addition to web and search history, modern web browsers contain a substantial amount of information that is not directly visible to the user.

The *browser cache* is a set of files that have been previously downloaded over the Internet. Browsers keep these duplicate copies of downloaded files in order to speed the web browsing experience: the cache eliminates the need to repeatedly download web objects such as decorative images or JavaScript functions that do not frequently change. The cache also provides a second history of the web user's actions. But there is no straightforward way in any of the browsers discussed in this chapter to visually inspect the cache and its contents, a violation of the EXPLICIT USER AUDIT design pattern. (The Netscape and Mozilla browsers implement a URL called `about:cache` that displays information about the files currently in the cache, but this URL appears to be relatively unknown; Internet Explorer similarly has a provision for viewing the folders that contain the cache files, but it is not obvious.)

Pages in the user's cache are deleted when they are not referenced for a period of time and new space is needed for new pages. But all three browsers have procedures for manually deleting the cached pages as well. One reason to delete these pages is when cached information is no longer valid, as can happen when a web site is under development. Users can also delete the pages in their browser cache when they are attempting to remove evidence from their computer that they have visited a particular web page.

Safari gives the user a straightforward control for clearing the contents of the cache: underneath the "Safari" application menu, there is a menu option labeled "Empty Cache..." Choosing this option displays a confirmatory alert panel which, if approved, causes the files in the cache to be unlinked.

Internet Explorer's control for clearing the cache is on the "General" tab of the "Internet Options" control panel. Microsoft uses different language from the other browsers—language that actually makes more sense but is nevertheless out-of-step with the other browsers. Instead of using the terminology "clear the cache" or something similar, the Internet Explorer command is labeled "Delete Files" and included in a box labeled "Temporary Internet Files" (Figure 4-6).

There are a variety of HCI-SEC problems that arise with this approach:

- Because the "History" view is disassociated from the pages in the cache, it is possible to clear the browser's history but still leave ample evidence that various web pages had been visited.
- Because the controls for deleting the cache are coarse-grained, a user's only realistic option for removing evidence that a web site was visited is to erase the entire browser cache. There are many circumstances in which such an action might generate suspicion.
- Because the browsers use different terminology and user interface elements, users must be specially trained to manage the cache for every web browser they use.

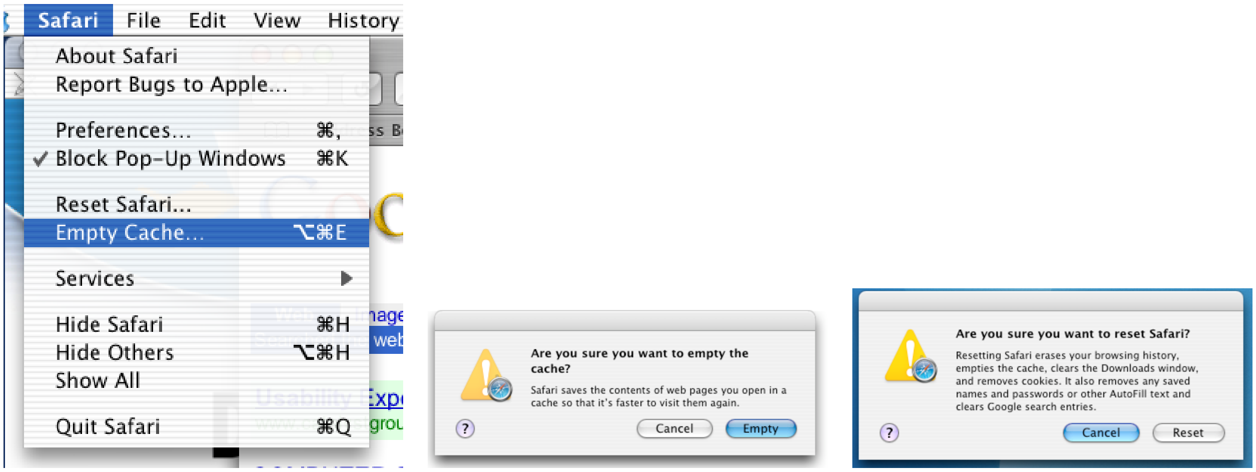
.

4.1.4 Implementing the Reset to Installation pattern

In addition to the personal information discussed in previous sections of this chapter, today's web browsers can store user-generated information in three other locations as well:

- Personal information that is used to automatically **fill in forms** on a web page.
- A **database of usernames and passwords** that have been memorized for web sites that require authentication.
- A **list of files that have been downloaded**, and the locations where they have been saved on the computer.

Apple Safari provides a simple way to remove all six types of personal information that can be captured in the browser: the "Reset Safari..." command, located on the program's main menu. Choosing this option causes Safari to delete the cache and all other personal information that Safari has accumulated—history, search history, cookies, bookmarks and so on—in one simple operation (Figure 4-8). (Unfortunately, the operation is confirmed with a pop-up menu, which provides protection against the command being accidentally chosen, but it does not implement the Delayed



Are you sure you want to empty the cache?

Safari saves the contents of web pages you open in a cache so that it's faster to visit them again.

Are you sure you want to reset Safari?

Resetting Safari erases your browsing history, empties the cache, clears the Downloads window, and removes cookies. It also removes any saved names and passwords or other AutoFill text and clears Google search entries.

Safari

Safari “Empty Cache...”

Safari “Reset Safari...”

Figure 4-8: Safari’s “File” menu has commands to “Empty Cache. . . ” and “Reset Safari. . . ” (left), which result in the warning panels (center and right, respectively) being displayed. Although emptying the cache is largely a non-destructive action, resetting the browser eliminates bookmarks and cookies, which can make it much harder (or even impossible) to access information on the web. The Safari browser doesn’t distinguish the severity of these two actions. Firefox has a similar buttons to individually clear history, saved password, the cache, or “all information stored while browsing,” as shown in Figure 4-5

UNRECOVERABLE ACTION pattern to cover mental slips.) This is an exact implementation of the RESET TO INSTALLATION pattern.

Firefox also implements the RESET TO INSTALLATION pattern, although it uses different terminology to implement the functionality, and the control is located in a different place. In Firefox the controls are located on the Privacy tab of the browser’s Preference Panel, (Figure 4-5), and the command is labeled “Clear all information stored while browsing.”

Once again, this confusion in both terminology and control placement detracts from usability, because it means that users who learn how to purge information in one browser cannot readily transfer that knowledge to an other. Security and usability could be aligned through the use of consistent terminology, as specified by the CONSISTENT MEANINGFUL VOCABULARY principle, and through the placement of the controls in consistent positions between the two browsers, as specified by the CONSISTENT CONTROLS AND PLACEMENT principle.

4.1.5 Solving the browser history usability problem with patterns

As developed in this chapter, the browser history problem arises because today’s web browsers do not implement the EXPLICIT USER AUDIT pattern. The fundamental problem is that web browsers retain information on the computer, but do not make this information visible to the computer user. Going deeper, we have shown browsers have failed to implement the EXPLICIT ITEM DELETE pattern: they frequently show information but do not give the user the ability to delete the information *where that information is shown*. And when browsers do give the user the ability to delete information, they do not delete it with COMPLETE DELETE. As a result, the information can be recovered using forensic means.

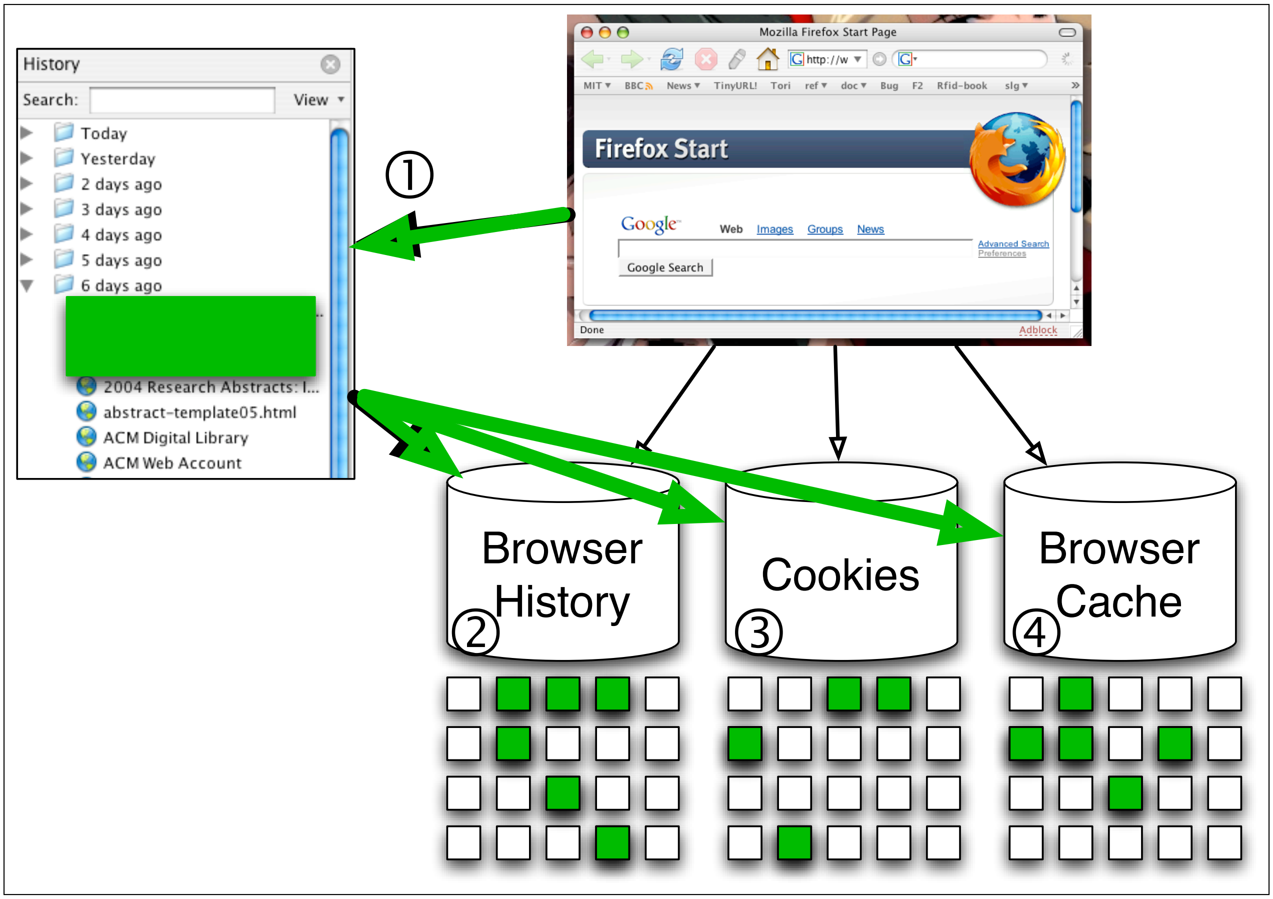
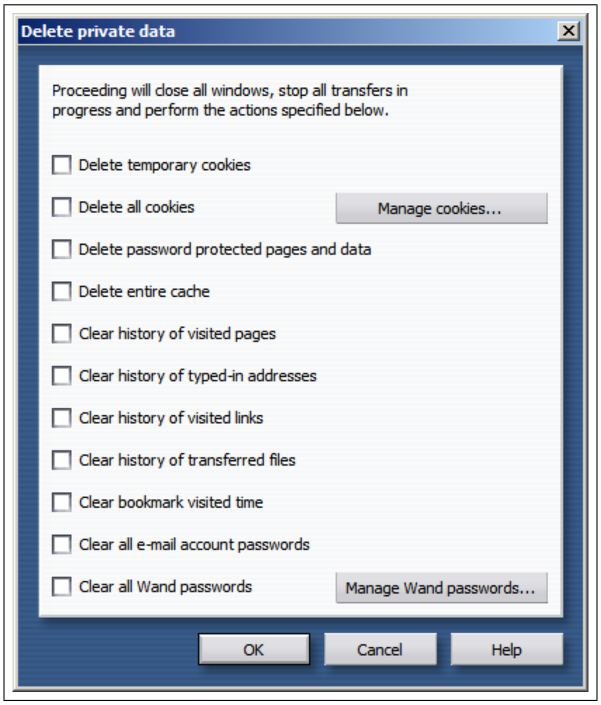


Figure 4-9: An illustration of the unified history and cache proposal. ①Deleting an element in the user-visible history should cause information to be deleted simultaneously in ②the browser history; ③the cookie jar; and ④the browser cache. If this deletion is accomplished with an overwriting delete, then the user can be assured that there will be no hidden history stored in the browser.

An alternative approach would be for browsers to implement the EXPLICIT ITEM DELETE pattern by giving the user the ability to delete the information where it is displayed, and implement the RESET TO INSTALLATION pattern, giving the user a simple way to eliminate *all* of the information that had been collected during the course of web browsing. In either case, deleting information from the browser's history should delete the matching information from the browser's cache and any cookies that pertain to the web site, as shown schematically in Figure 4-9.

It makes sense to delete the cookies if the user is explicitly trying to delete evidence that a web site was visited. If the cookies are not deleted, the cookies constitute hidden evidence that a web site was visited. Likewise, the patterns suggest that references in the browser's history should not be deleted if the computer contains a persistent cookie: otherwise the computer will be violating the EXPLICIT USER AUDIT pattern. Indeed, much of the early outrage over cookies in 1996 and 1997 was due to the fact that tracking cookies had been placed on user computers without the permission.[Gar96a]



Proceeding will close all windows, stop all transfers in progress and perform the actions specified below.

□ Delete temporary cookies
□ Delete all cookies
□ Delete password protected pages and data
□ Delete entire cache
□ Clear history of visited pages
□ Clear history of typed-in addresses
□ Clear history of visited links
□ Clear history of transferred files
□ Clear bookmark visited time
□ Clear all e-mail account passwords
□ Clear all Wand passwords

Figure 4-10: The Opera web browser has a command called "Delete Private Data," which displays a panel that gives the user a great deal of control over what kind of private data is actually deleted.

Why is there no Complete Delete?

As part of the work performed in this chapter, attempts were made to determine why major web browsers do not implement the Clean Delete design pattern. While none of the major browser vendors provided an explanation for the lack of functionality, Opera Software's Chief Technology Officer, Håkon Wium Lie, was willing to explain why Complete Delete is not implemented in the Opera web browser.

User privacy has always been a developer concern at Opera Software. Indeed, the Opera browser was the first browser to implement the Reset to Installation pattern with a "Delete Private Data" command (Figure 4-10). But while this command gives the user a great deal of control over the types of private data deleted—including cookies, passwords, the cache, and other information—the browser does not use Complete Delete to actually perform the deletion. Instead, the information is left behind on the disk!

Opera could reduce the risk that this data would be recovered at a later point in time by explicitly overwriting the files before they were deleted. But according to Lie, there was a formal decision made not to implement such functionality:

> "The problem is that it's hard—if not impossible—to guarantee that bits will disappear from the disk. In normal operation, files grow and shrink and bits will be left here and there. When Opera is told to 'Delete Private Data' (which I think is a unique and valuable tool), we could overwrite the current file, but there may still be bits lying around from recent shrinks.

> “Also, with the advent of journaling file systems, the OS will retain information even after the application ’overwrite’ the bits.
>
> “So, to conclude, we have no way of guaranteeing that the bits disappear. If you need security at that level, it’s probably best to use a specialized file system tool in combination with Opera.”[Lie04]

This is a common sentiment among security practitioners. They fear that some users may be misled and come to rely on that incomplete solution. It is better, these practitioners argue, to provide no solution at all, than to provide a solution that offers incomplete security. But this argument is flawed especially here: the browser is *already* misleading users by making it appear that data has been deleted, when in fact that data has merely been made invisible. Even a partial implementation of COMPLETE DELETE—for example, by explicitly overwriting the files before unlinking them—would be better than no solution, since the partial solution would leave sensitive information on the computer’s hard drive. The partial solution might still mislead some people some of the time, but it would almost certainly mislead most people less frequently.

Lie’s viewpoint, in fact is the reason that this thesis proposes the principle DEPLOY GOOD SECURITY (DON’T WAIT FOR PERFECT).

As an aside, the large number of sanitization options provided by Opera’s “Delete Private Data” panel is a violation of the PROVIDE STANDARDIZED SECURITY POLICIES principle. It is unlikely that most of Opera’s users understand the security implications of deleting vs. not deleting a specific class of information. An alternative approach would be to provide a default deletion policy—delete all of the data—and then allow this to be customized through the use of an “advanced” button if necessary. It would also be useful if this functionality were implemented with the DELAYED UNRECOVERABLE ACTION pattern, so that the user could experience web browsing without the private data, prior to having the private data being irrevocably erased.

4.1.6 Consumer education: the anti-pattern

Instead of fixing these fundamental problems in web browsers, both Microsoft and Internet service providers have spent considerable effort on educating users about the importance of deleting confidential information from the browser’s cache and history.[Mic03a, Com03] Yet the instructions that these organizations give are frequently incomplete. For example, [Mic03a] explains how to clear Internet Explorer’s history and cached addresses in the Address box. But [Mic03a] does not explain how to clear the browser cache—that is explained on another Microsoft web page [Mic04], and this second web page does not mention how to clear the browser’s history or the Address box. Worse, there is no linkage between these two pages. Furthermore, the instructions in [Mic03a] require manually deleting a Registry key—a procedure that [Mic03a] does not explain.

Some organizations—even very small ones—have taken matters into their own hands. For example, HopeForHealing.org, a small web site devoted to helping the survivors of sexual and domestic abuse, devotes considerable information on the home page of its web site to instructions on how to clear the browser’s history and cache.[Hop04] “Click here to learn how to clear your browser’s history if visiting this page puts you at risk,” reads a banner link across the site’s home page, with a link to detailed instructions on how to erase the cache and history of both Internet Explorer and

Netscape Navigator. The web page further suggests that it is good practice for women who are in danger to clear the "redial" button on touch-tone telephones after calling a shelter!

In November 2004, a Google search for web pages that contained the phrase "Internet options" and "Clear history" returned 17,400 sites; by March 2005 the number of web sites had risen to 20,400, indicating that a growing number of organizations believe they must educate users regarding this browser arcana. But a better approach would be to fix the underlying paradigm that causes the browser's stored data to be inconsistent with the view that is provided to the user.

4.1.7 Future work

The information presented in this section is based on 12 years' of personal experience with web browsers and an evaluation of web browser sanitization practices that has lasted for at least two years. The logical extension of this work would be to conduct further user studies and surveys to determine whether or not the conclusions reached in this section apply to more mainstream users.

Specifically, user studies could be carried out to determine if typical computer users are aware of the facilities included in web browsers for removing traces of web activity. Apple's technique of putting the "Clear History" command at the bottom of the History menu should be formally evaluated to see if this really is an approach that could be broadly applied, or if it unacceptably increases the chances of accidently clearing a user's history.

It should be possible to modify the open-source Firefox web browser to see if the link between browser history and cache is feasible. Likewise, it would be interesting to modify Firefox to evaluate the performance impact of a sanitizing delete and to evaluate techniques for removing the performance penalty.

Finally, it may be useful to evaluate add-on software that is currently providing sanitization services to see if these programs actually do what they say. Geiger's initial investigation finds them lacking.[Gei04]

4.2 Case Study: Failed Document Sanitization in Word and Acrobat

With the growth of the Web as a means for publishing documents in the 1990s, there have been a significant number of incidents in which confidential information—and occasionally US Government classified information—was inadvertently released in Adobe Acrobat and Microsoft Word documents. Once again, the problem is that hidden information that could not be audited or deleted—this time in the Acrobat and Word file format. That is, Adobe Acrobat and Microsoft Word do not follow the EXPLICIT USER AUDIT and COMPLETE DELETE design patterns.

4.2.1 Media reports

In recent years there have been several cases in which confidential information was revealed as a result of organizations posting documents on the Internet containing hidden information, after which the documents were downloaded and the information revealed by others. There has also at least one high-profile case in which an organization resorted to scanning a redacted document and placing the scan on the Internet. By scanning the document, the organization created a kind of "optical firewall" that prevented hidden information in the electronic document from leaking into the Acrobat scan.

- ***New York Times*, June 2000:** After obtaining a classified CIA file documenting how American and British officials engineered the 1953 coup that overthrew Iran's elected government, editors at *The New York Times* decided to put the file on the newspaper's web site. In order to protect the identities of the two dozen Iranians whose name appeared in the document, the *Times* placed black boxes over the names, for fear that publishing the names might place the individuals or their families at risk. After the file was posted, John Young, editor of CRYPTOME, downloaded the file and viewed it on a very slow computer. Young noticed that the Adobe Acrobat software was actually displaying the names and then covering them over with a black boxes! Young contacted the newspaper and was asked to keep the names confidential, but Young decided to publish them on his web site.[Won00]

- **US Department of Justice, October 2003:** When the US Justice Department released its June 2002 Workplace Diversity report in October 2003, the version of the report that was placed on the Department's web site had been heavily edited to delete criticisms of the Department's policies. The "editing" was in the form of black boxes that had been placed over the embarrassing text. Journalists were able to remove the black boxes and disclose the embarrassing information.[Edm03, Joh04] Later the MemoryHole.Org web site placed an unredacted version of the report on its web site.[Pou03, Kic03]

- **SCO Group, March 2004:** When the SCO Group filed lawsuits against DaimlerChrysler and AutoZone for using Linux, an analysis of the Microsoft Word files conducted by journalists revealed that SCO had previously planned to target Bank of America.[2][SA04]

- **Multinational Forces-Iraq (MNF-I) report on the death of Nicola Calipari, April 2005:** After the mistaken killing of an Italian intelligence agent on March 4th, 2005 in Iraq, the

[2] According to the Shankland and Ard, "on Feb. 18 at 11:10 a.m. 'Bank of America, a National Banking Association' was removed as a defendant and 'DaimlerChrysler Corp.' was inserted. Three minutes later, this comment was removed: 'Are there any special jurisdiction or venue requirements for a NA bank?' " Delete comments were also found in the document, such as "Did BA receive one of the SCO letters sent to Fortune 1500?"[SA04]

Multinational Forces-Iraq (MNF-I) overseen by the United States military performed an internal investigation regarding the circumstances of the killing. A redacted report was uploaded to a US Department of Defense web site on April 30th, 2005.[CNN05]

The report had been redacted by drawing black boxes over the pages of the Adobe Acrobat file, leaving the original text underneath the boxes. Two days later, a German systems architect named Volker Weber was able to recover the entire text of the document with two keystrokes: by selecting all of the by typing control-A, and then copying all of the text with control-C.[Ber05a]

The redaction can be shown visually through the use of Adobe Illustrator CS, which has the ability to directly edit PDF files and remove the redacting boxes, as shown in Figure 4-11.

- **Byers: 10% of Microsoft Word files on the Internet have substantial hidden content.** In 2003 Simon Byers, an AT&T researcher, downloaded 100,000 Word documents from web sites located all over the Internet. He then examined the files using an automated technique and determined that approximately half of the documents he downloaded contained between 10 and 50 hidden words, a third had between 50 and 500 words, and 10% had more than 500 words. [Bye03]

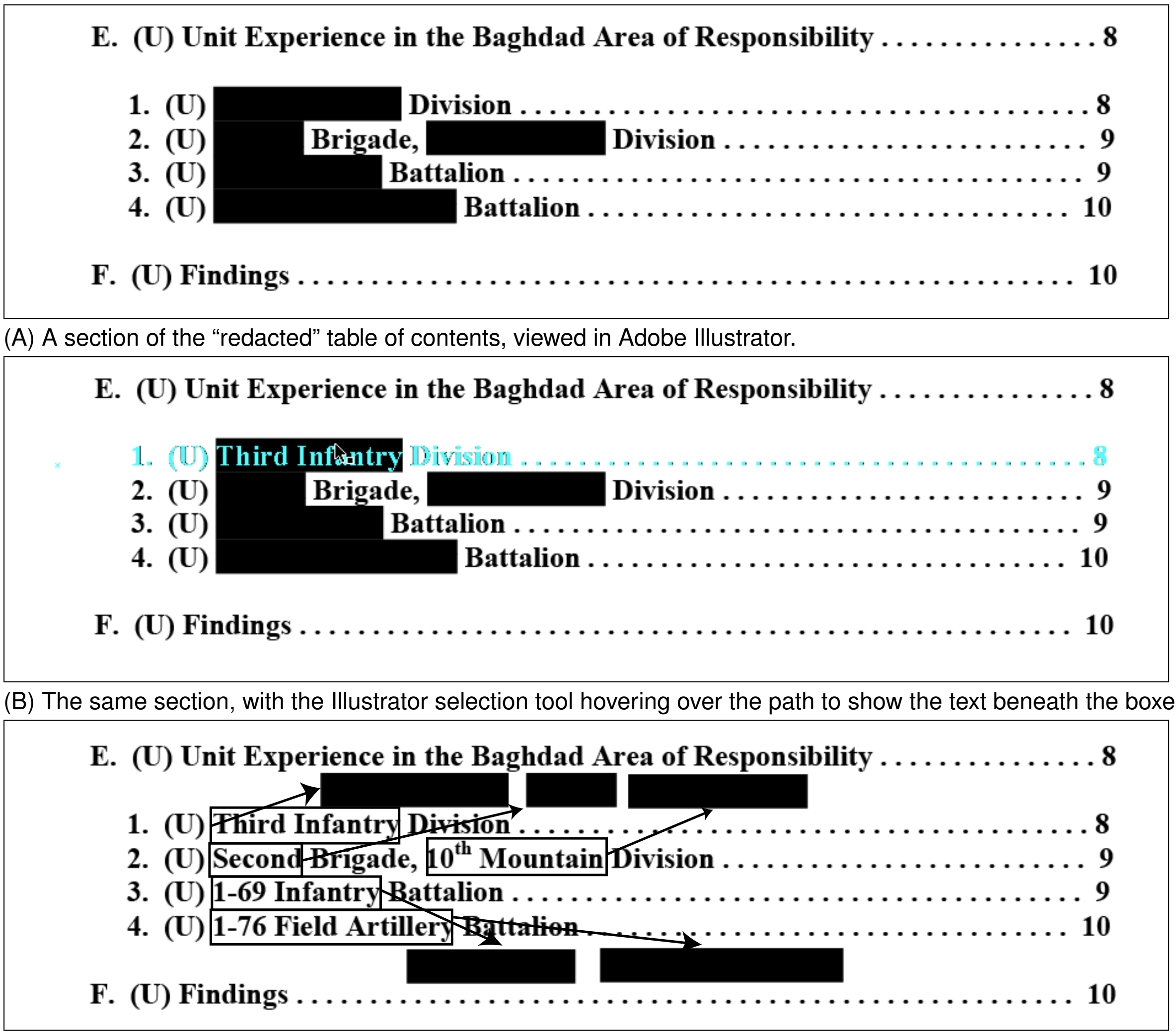
At least some organizations appear to be aware of the risk of hidden information in documents. After concluding a classified investigation into the intelligence failures leading up to the US war with Iraq in 2003, the US Senate Intelligence Committee issued a “Report on the US Intelligence Community’s Prewar Intelligence Assessments on Iraq.” The report was published on July 10th, 2004, as an Adobe Acrobat file on the Committee’s web site. But instead of publishing an Acrobat file that contained text, the Committee’s published Acrobat file contained page after page of scanned images that clearly had been eradicated after printing and before they were scanned. Whereas an Acrobat file of just the text would have been only a few megabytes, the Acrobat image file was over 13 megabytes.

By producing an Acrobat file from a scan of a printout of the sanitized document, the Committee ensure that no hidden information in the original document would leak from the original document into the final Acrobat file that was placed on the Internet. The original document contained many instances of security classification labels at the beginning of paragraphs—an “(S)” symbol indicating that a paragraph contained secret information, and a “(TS)” symbol indicating that the paragraph contained top secret information. Given the value of the sanitized information, this trip from the electronic realm into the optical realm—a kind of “optical firewall”—might well have been appropriate. Unfortunately, the publication of images instead of text was a clear violation of spirit of Section 508 of the Rehabilitation Act, since the scanned images could not be processed by a screen reader. (Of course, as Section 508 is a procurement regulation, it does not apply to the US Senate Intelligence Committee. For more information on Section 508, see Section 2.6.6.)

Despite repeated repeated requests, the Committee refused to comment as to why the report was prepared in this manner.

4.2.2 Analysis of Microsoft Word

While hidden content has been found in both documents created with Microsoft Word and Adobe Acrobat, the causes of problems on those two platforms in quite different. With Microsoft Word, the



E. (U) Unit Experience in the Baghdad Area of Responsibility 8

1. (U) [redacted] Division 8
2. (U) [redacted] Brigade, [redacted] Division 9
3. (U) [redacted] Battalion 9
4. (U) [redacted] Battalion 10

F. (U) Findings 10

(A) A section of the "redacted" table of contents, viewed in Adobe Illustrator.

E. (U) Unit Experience in the Baghdad Area of Responsibility 8

1. (U) Third Infantry Division 8
2. (U) [redacted] Brigade, [redacted] Division 9
3. (U) [redacted] Battalion 9
4. (U) [redacted] Battalion 10

F. (U) Findings 10

(B) The same section, with the Illustrator selection tool hovering over the path to show the text beneath the boxes.

E. (U) Unit Experience in the Baghdad Area of Responsibility 8

1. (U) Third Infantry Division 8
2. (U) Second Brigade, 10th Mountain Division 9
3. (U) 1-69 Infantry Battalion 9
4. (U) 1-76 Field Artillery Battalion 10

F. (U) Findings 10

(C) The same section, with the black boxes moved aside, revealing the classified headings.

Figure 4-11: Using Adobe Illustrator to un-redact a section of the Multinational Forces-Iraq (MNF-I) report on the death of Italian intelligence agent Nicola Calipari.

problem is caused by a combination of the Word file format and the program's "fast saves" feature; problems are also caused by Word's facilities for revision and change tracking.

Designed when computers were much slower and had less memory than today, the Word file format is largely a dump of the application's memory followed by a series of changes that are to be applied to the memory image after the document is loaded. This format allowed Microsoft to implement a "fast save" feature, in which a few minor changes to a document could simply be appended to the end of the document file. This made it possible to open a 100-page document, make a few changes, and save it out again within a matter of seconds—even if the document was many times larger than the computer's available memory.

A result of this "fast save" feature is that the Word document file might contain information that was intentionally removed by the operator. For this reason modern versions of Word require that the "fast save" feature be explicitly enabled, as shown in Figure 4-12. Other Microsoft Office programs, including PowerPoint and Excel, have similar "fast save" features.

Microsoft Word also has extensive provisions for tracking revisions, author information, comments, and even for checkpointing complete documents. All of these features store metadata in the Word file format. Experience has shown that many Word users are not aware of the extent of information that is captured.

As discussed in Section 2.5.6, Microsoft has created a "Remove Hidden Data" tool that will remove the hidden information from Word files. But it is unlikely that organizations even know that the tool exists, let alone have trained their employees in its use. Finally, as Byers notes, there is no easy way to look at a Microsoft Word file and determine if the hidden data has been deleted or not.

Byers recommended that organizations not use Microsoft Word files as a publication format for external web sites.[Bye03] Unfortunately, this recommendation isn't workable: many employees simply do not have the training to convert documents into other file formats, and often there is a desire to make documents available in editable form.

4.2.3 Analysis of Adobe Acrobat

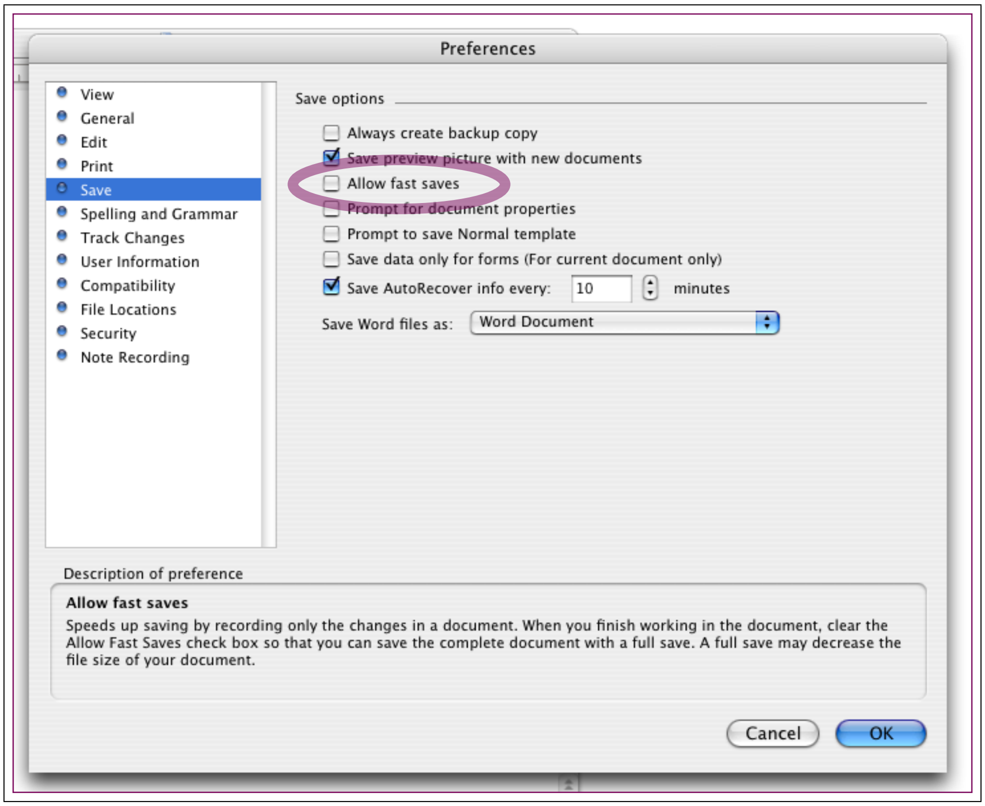
The disclosure of the data resulting from the improper use of the Acrobat draw-box feature deserves special mention. Placing black boxes over confidential or classified information and then photocopying the documents has been a standard way to eradicate such information from documents for decades. It is not surprising, then, that drawn black boxes might be used by untrained individuals for the purpose of eradication.

Ironically, there is a plug-in for Adobe Acrobat called Redax that allows users to still use this intuitive metaphor, but Redax which actually erases the information that is covered-up.[App03] The tool, when loaded into Adobe Acrobat, causes the combined system to implement the EXPLICIT ITEM DELETE pattern.

Redax is designed for use by federal agencies that need to comply with Freedom of Information Act (FOIA) requests without forcing them to print, redact, and then re-scan documents that they wish to distribute in electronic form. Redax also supports the insertion of FOIA "Exemption codes" which are used to indicate in a systematic matter the FOIA exemption that was used to justify the redaction. The plug-in features an interface that lets a government information officer mark with a black box the areas of the document that are to be redacted—a metaphor that is similar to the black magic markers employed by most censors. But rather than covering the information with a black square, Redax actually removes the information from the underlying document. The program also replaces the "text" with hyphens so that exported text will clearly indicate that a redaction has taken place.

4.3 Conclusion

This chapter has shown that there are many cases in which potentially confidential information is present but not visible in the databases maintained by web browsers and in the document files pro-



Save options

□ Always create backup copy

☑ Save preview picture with new documents

□ Allow fast saves

□ Prompt for document properties

□ Prompt to save Normal template

□ Save data only for forms (For current document only)

☑ Save AutoRecover info every: 10 minutes

Save Word files as: Word Document

Description of preference

Allow fast saves

Speeds up saving by recording only the changes in a document. When you finish working in the document, clear the Allow Fast Saves check box so that you can save the complete document with a full save. A full save may decrease the file size of your document.

Figure 4-12: The Preferences panel of Microsoft Word has an option labeled "Allow fast saves." The in-program documentation explains that allowing fast saves will shorten the time required to save large documents. Fast saves work by appending user changes to the document as a series of transactions to the end of the document file. Fast saves can also silently compromise privacy or security by leaving confidential information in the document file after the user has intentionally tried to eliminate that information. Unfortunately, this aspect of fast saves is not addressed by the in-program documentation.

duced by Microsoft Word and Adobe Acrobat. We have also seen that the same patterns introduced in Chapter 3 to cover disk and file system sanitization issues can be used here to cover sanitization issues in a different domain. These patterns will be fully described in Chapter 10.