CHAPTER 3

Ghosts in the Machine

Secrets and Surprises of Electronic Documents

What You See Is Not What the Computer Knows

On March 4, 2005, Italian journalist Giuliana Sgrena was released from captivity in Baghdad, where she had been held hostage for a month. As the car conveying her to safety approached a checkpoint, it was struck with gunfire from American soldiers. The shots wounded Sgrena and her driver and killed an Italian intelligence agent, Nicola Calipari, who had helped engineer her release.

A fierce dispute ensued about why U.S soldiers had rained gunfire on a car carrying citizens of one of its Iraq war allies. The Americans claimed that the car was speeding and did not slow when warned. The Italians denied both claims. The issue caused diplomatic tension between the U.S. and Italy and was a significant political problem for the Italian prime minister.

The U.S. produced a 42-page report on the incident, exonerating the U.S. soldiers. The report enraged Italian officials. The Italians quickly released their own report, which differed from the U.S. report in crucial details.

Because the U.S. report included sensitive military information, it was heavily redacted before being shared outside military circles (see Figure 3.1). In another time, passages would have been blacked out with a felt marker, and the document would have been photocopied and given to reporters. But in the information age, the document was redacted and distributed electronically, not physically. The redacted report was posted on a web site the allies used to provide war information to the media. In an instant, it was visible to any of the world's hundreds of millions of Internet users.

(U) [redacted] has Direct Liaison Authorized (DIRLAUTH) to coordinate directly with [redacted] for security along Route Irish. This is the same level of coordination previously authorized by [redacted] Division to [redacted]. When executing DIRLAUTH, [redacted] directly coordinates an action with units internal or external to its command and keeps the [redacted] commander informed. The [redacted] TOC passes all coordination efforts through the [redacted] Brigade TOC to [redacted] JOC. (Annex 58C).

Source: http://www.corriere.it/Media/Documenti/Classified.pdf, extract from page 10.

FIGURE 3.1 Section from page 10 of redacted U.S. report on the death of Italian journalist Nicola Calipari. Information that might have been useful to the enemy was blacked out.

One of those Internet users was an Italian blogger, who scrutinized the U.S. report and quickly recovered the redacted text using ordinary office software. The blogger posted the full text of the report (see Figure 3.2) on his own web site. The unredacted text disclosed positions of troops and equipment, rules of engagement, procedures followed by allied troops, and other information of interest to the enemy. The revelations were both dangerous to U.S. soldiers and acutely embarrassing to the U.S. government, at a moment when tempers were high among Italian and U.S. officials. In the middle of the most high-tech war in history, how could this fiasco have happened?

(U) 1-76 FA has Direct Liaison Authorized (DIRLAUTH) to coordinate directly with 1-69 IN for security along Route Irish. This is the same level of coordination previously authorized by 1st Cavalry Division to 2-82 FA. When executing DIRLAUTH, 1-76 FA directly coordinates an action with units internal or external to its command and keeps the 31D commander informed. The 1-76 FA TOC passes all coordination efforts through the 4th Brigade TOC to 31D JOC. (Annex 58C).

Source: http://www.corriere.it/Media/Documenti/Unclassified.doc.

FIGURE 3.2 The text of Figure 3.1 with the redaction bars electronically removed.

Paper documents and electronic documents are useful in many of the same ways. Both can be inspected, copied, and stored. But they are not equally useful for all purposes. Electronic documents are easier to change, but paper documents are easier to read in the bathtub. In fact, the metaphor of a series of bits as a "document" can be taken only so far. When stretched beyond its breaking point, the "document" metaphor can produce surprising and damaging results—as happened with the Calipari report.

Office workers love "WYSIWYG" interfaces—"What You See Is What You Get." They edit the electronic document on the screen, and when they print it, it looks just the same. They are deceived into thinking that what is in the

computer is a sort of miniaturized duplicate of the image on the screen, instead of computer codes that produce the picture on the screen. In fact, the WYSIWYG metaphor is imperfect, and therefore risky. The report on the death of Nicola Calipari illustrates what can go wrong when users accept such a metaphor too literally. What the authors of the document saw was dramatically different from what they got.

The report had been prepared using software that creates PDF files. Such software often includes a "Highlighter Tool," meant to mimic the felt markers that leave a pale mark on ordinary paper, through which the underlying text is visible (see Figure 3.3). The software interface shows the tool's icon as a marker writing a yellow stripe, but the user can change the color of the stripe. Probably someone tried to turn the Highlighter Tool into a redaction tool by changing its color to black, unaware that what was visible on the screen was not the same as the contents of the electronic document.

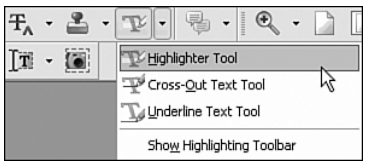
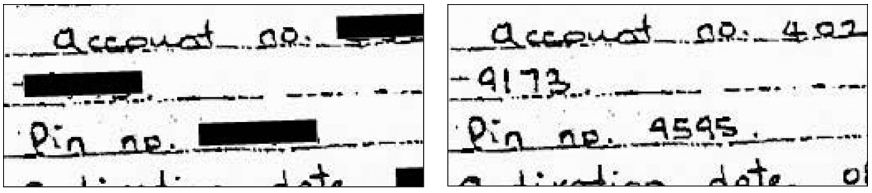


FIGURE 3.3 Adobe Acrobat Highlighter Tool, just above the middle. On the screen, the "highlighter" is writing yellow ink, but with a menu command, it can be changed to any other color.

The Italian blogger guessed that the black bars were nothing more than overlays created using the Highlighter Tool, and that the ghostly traces of the invisible words were still part of the electronic document that was posted on the web. With that realization, he easily undid the black "highlighting" to reveal the text beneath.

Just as disturbing as this mistake is the fact that two major newspapers had quite publicly made the same mistake only a few years before. On April 16, 2000, the *New York Times* had detailed a secret CIA history of attempts by the U.S. to overthrow Iran's government in 1953. The newspaper reproduced sections of the CIA report, with black redaction bars to obscure the names of CIA operatives within Iran. The article was posted on the Web in mid-June, 2000, accompanied by PDFs of several pages of the CIA report. John Young, who administers a web site devoted to publishing government-restricted documents, removed the redaction bars and revealed the names of CIA agents. A controversy ensued about the ethics and legality of the disclosure, but the names are still available on the Web as of this writing.

The *Washington Post* made exactly the same mistake in 2002, when it published an article about a demand letter left by the Washington snipers, John Allen Muhammad and John Lee Malvo. As posted on the *Post*'s web site, certain information was redacted in a way that was easily reversed by an inquisitive reader of the online edition of the paper (see Figure 3.4). The paper fixed the problem quickly after its discovery, but not quickly enough to prevent copies from being saved.



Source: Washington Post web site, transferred to web.bham.ac.uk/forensic/news/02/sniper2.html. Actual images taken from slide 29 of http://www.ccc.de/congress/2004/fahrplan/files/316-hidden-data-slides.pdf.

FIGURE 3.4 Letter from the Washington snipers. On the left, the redacted letter as posted on the *Washington Post* web site. On the right, the letter with the redaction bars electronically removed.

What might have been done in these cases, instead of posting the PDF with the redacted text hidden but discoverable? The Adobe Acrobat software has a security feature, which uses encryption (discussed in Chapter 5, "Secret Bits") to make it impossible for documents to be altered by unauthorized persons, while still enabling anyone to view them. Probably those who created these documents did not know about this feature, or about commercially available software called Redax, which government agencies use to redact text from documents created by Adobe Acrobat.

A clumsier, but effective, option would be to scan the printed page, complete with its redaction bars. The resulting file would record only a series of black and white dots, losing all the underlying typographical structure—font names and margins, for example. Whatever letters had once been "hidden" under the redaction bars could certainly not be recovered, yet this solution has an important disadvantage.

One of the merits of formatted text documents such as PDFs is that they can be "read" by a computer. They can be searched, and the text they contain can be copied. With the document reduced to a mass of black and white dots, it could no longer be manipulated as text.

A more important capability would be lost as well. The report would be unusable by programs that vocalize documents for visually impaired readers. A blind reader could "read" the U.S. report on the Calipari incident, because software is available that "speaks" the contents of PDF documents. A blind reader would find a scanned version of the same document useless.

Tracking Changes—and Forgetting That They Are Remembered

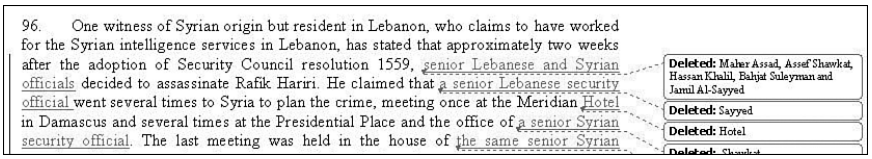
In October, 2005, UN prosecutor Detlev Mehlis released to the media a report on the assassination of former Lebanese Prime Minister Rafik Hariri. Syria had been suspected of engineering the killing, but Syrian President Bashar al-Assad denied any involvement. The report was not final, Mehlis said, but there was "evidence of both Lebanese and Syrian involvement." Deleted, and yet uncovered by the reporters who were given the document, was an incendiary claim: that Assad's brother Maher, commander of the Republican Guard, was personally involved in the assassination.

Microsoft Word offers a "Track Changes" option. If enabled, every change made to the document is logged as part of the document itself—but ordinarily not shown. The document bears its entire creation history: who made each change, when, and what it was. Those editing the document can also add comments—which would not appear in the final document, but may help editors explain their thinking to their colleagues as the document moves around electronically within an office.

Of course, information about strategic planning is not meant for outsiders to see, and in the case of legal documents, can have catastrophic consequences if revealed. It is a simple matter to remove these notes about the document's history—but someone has to remember to do it! The UN prosecutor neglected to remove the change history from his Microsoft Word document, and a reporter discovered the deleted text (see Figure 3.5). (Of course, in Middle Eastern affairs, one cannot be too suspicious. Some thought that Mehlis had intentionally left the text in the document, as a warning to the Syrians that he knew more than he was yet prepared to acknowledge.)

A particularly negligent example of document editing involved SCO Corporation, which claimed that several corporations violated its intellectual property rights. In early 2004, SCO filed suit in a Michigan court against Daimler Chrysler, claiming Daimler had violated terms of its Unix software agreement with SCO. But the electronic version of its complaint carried its modification history with it, revealing a great deal of information about SCO's litigation planning. In particular, when the change history was revealed, it

turned out that until exactly 11:10 a.m. on February 18, 2004, SCO had instead planned to sue a different company, Bank of America, in federal rather than state court, for copyright infringement rather than breach of contract!


96. One witness of Syrian origin but resident in Lebanon, who claims to have worked for the Syrian intelligence services in Lebanon, has stated that approximately two weeks after the adoption of Security Council resolution 1559, senior Lebanese and Syrian officials decided to assassinate Rafik Hariri. He claimed that a senior Lebanese security official went several times to Syria to plan the crime, meeting once at the Meridian Hotel in Damascus and several times at the Presidential Place and the office of a senior Syrian security official. The last meeting was held in the house of the same senior Syrian

Deleted: Maher Assad, Assef Shawkat, Hassan Khalil, Bahjat Suleyman and Jamil Al-Sayyed

Deleted: Sayyed

Deleted: Hotel

Source: Section of UN report, posted on Washington Post web site, `www.washingtonpost.com/wp-srv/world/syria/mehlis.report.doc`.

FIGURE 3.5 Section from the UN report on the assassination of Rafik Hariri. An earlier draft stated that Maher Assad and others were suspected of involvement in the killing, but in the document as it was released, their names were replaced with the phrase "senior Lebanese and Syrian officials."

Saved Information About a Document

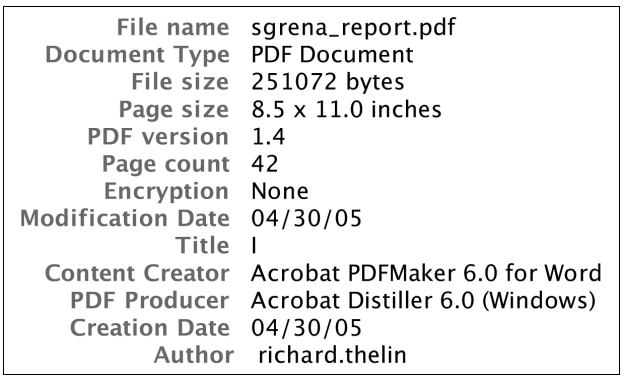
FORGING METADATA

Metadata can help prove or refute claims. Suppose Sam emails his teacher a homework paper after the due date, with a plea that the work had been completed by the deadline, but was undeliverable due to a network failure. If Sam is a cheater, he could be exposed if he doesn't realize that the "last modified" date is part of the document. However, if Sam *is* aware of this, he could "stamp" the document with the right time by re-setting the computer's clock before saving the file. The name in which the computer is registered and other metadata are also forgeable, and therefore are of limited use as evidence in court cases.

An electronic document (for example, one produced by text-processing software) often includes information that is *about* the document—so-called *metadata*. The most obvious example is the name of the file itself. File names carry few risks. For example, when we send someone a file as an email attachment, we realize that the recipient is going to see the name of the file as well as its contents.

But the file is often tagged with much more information than just its name. The metadata generally includes the name associated with the owner of the computer, and the dates the file was created and last modified—often useful information, since the recipient can tell whether she is receiving an older or newer version than the version she already

has. Some word processors include version information as well, a record of who changed what, when, and why. But the unaware can be trapped even by such innocent information, since it tends not to be visible unless the recipient asks to see it. In Figure 3.6, the metadata reveals the name of the military officer who created the redacted report on the death of Nicola Calipari.



File name	sgrena_report.pdf
Document Type	PDF Document
File size	251072 bytes
Page size	8.5 x 11.0 inches
PDF version	1.4
Page count	42
Encryption	None
Modification Date	04/30/05
Title	I
Content Creator	Acrobat PDFMaker 6.0 for Word
PDF Producer	Acrobat Distiller 6.0 (Windows)
Creation Date	04/30/05
Author	richard.thelin

FIGURE 3.6 Part of the metadata of the Calipari report, as revealed by the "Properties" command of Adobe Acrobat Reader. The data shows that Richard Thelin was the author, and that he altered the file less than two minutes after creating it. Thelin was a Lieutenant Colonel in the U.S. Marine Corps at the time of the incident.

Authorship information leaked in this way can have real consequences. In 2003, the British government of Tony Blair released documentation of its case for joining the U.S. war effort in Iraq. The document had many problems—large parts of it turned out to have been plagiarized from a 13-year-old PhD thesis. Equally embarrassing was that the electronic fingerprints of four civil servants who created it were left on the document when it was released electronically on the No. 10 Downing Street web site. According to the *Evening Standard of London*, "All worked in propaganda units controlled by Alastair Campbell, Tony Blair's director of strategy and communications," although the report had supposedly been the work of the Foreign Office. The case of the "dodgy dossier" caused an uproar in Parliament.

You don't have to be a businessperson or government official to be victimized by documents bearing fingerprints. When you send someone a document as an attachment to an email, very likely the document's metadata shows who actually created it, and when. If you received it from someone else

and then altered it, that may show as well. If you put the text of the document into the body of your email instead, the metadata won't be included; the message will be just the text you see on the screen. Be sure of what you are sending before you send it!

Can the Leaks Be Stopped?

Even in the most professional organizations, and certainly in ordinary households, knowledge about technological dangers and risks does not spread instantaneously to everyone who should know it. The Calipari report was published five years after the *New York Times* had been embarrassed. How can users of modern information technology—today, almost all literate people—stay abreast of knowledge about when and how to protect their information?

It is not easy to prevent the leakage of sensitive information that is hidden in documents but forgotten by their creators, or that is captured as metadata. In principle, offices should have a check-out protocol so that documents are cleansed before release. But in a networked world, where email is a critical utility, how can offices enforce document release protocols without rendering simple tasks cumbersome? A rather harsh measure is to prohibit use of software that retains such information; that was the solution adopted by the British government in the aftermath of the "dodgy dossier" scandal. But the useful features of the software are then lost at the same time. A protocol can be established for converting "rich" document formats such as that of Microsoft Word to formats that retain less information, such as Adobe PDF. But it turns out that measures used to eradicate personally identifiable information from documents don't achieve as thorough a cleansing as is commonly assumed.

At a minimum, office workers need education. Their software has great capabilities they may find useful, but many of those useful features have risks as well. And we all just need to think about what we are doing with our documents. We all too mindlessly re-type keystrokes we have typed a hundred times in the past, not pausing to think that the hundred and first situation may be different in some critical way!

Representation, Reality, and Illusion

René Magritte, in his famous painting of a pipe, said "This isn't a pipe" (see Figure 3.7). Of course it isn't; it's a painting of a pipe. The image is made out