

What your documents say about you

The page is what you checked. The file is what you sent, and it carries a second document nobody read

For anyone who sends client documents out of the firm, and whoever wrote the policy they are following | 08 September 2026

A file is not a picture of a page. It is a container, and alongside the words you proofread it holds who wrote them, on whose machine, what the file was called before it was renamed, what a summary figure was calculated from, and — in a deck — what the presenter was told to say and not to show.

None of it is on the page, which is why nobody checks it. This paper sets out what each format carries, why removing it is more than one operation, and the two decisions worth refusing to automate.

1. The second document

The review process in most firms is aimed at the page. Someone reads the words, checks the figures, and signs it off. That review is real work and it is not the thing that goes wrong.

What leaves the building is the file, and the file contains a second document that no part of that review touched. It was written automatically, by the applications the work passed through, and it accumulated quietly over every save.

The failures that follow have a distinctive shape. Nobody was careless. A document was prepared properly, reviewed properly, and sent — and something the sender did not know was in it turned out to be the story: an author's name on a document meant to be from the firm rather than a person, an earlier filename that revealed the deal it was recycled from, a summary figure with its workings still attached, a slide deck carrying the note about what not to mention.

The reason this survives every training programme is that there is nothing to notice. The document looks exactly the way it should. What is missing from the reader's view is missing from the sender's too.

2. What is actually in there

By format, because the containers are genuinely different and a habit that works for one does nothing for another.

FORMAT	WHAT IT CARRIES BEYOND THE PAGE	WHY IT IS THERE
PDF	An information dictionary naming the author, the producing application and often the title the file had before it was renamed. Separately, an XMP stream keeping a history of the same. Possibly embedded scripting, actions that run on open, and attached files.	Every producing application writes it. The dictionary is not a fixed schema, so a producer may put anything at all in it.
Spreadsheets	Creator, last modified by, title, subject, description, keywords, category, when it was last printed, company and manager — plus a custom properties part that can hold whatever a template or a document management system decided to attach.	Written by the application, and often extended by whatever system the file passed through on its way.
Presentations	Speaker notes, one part per slide, sitting in the file whether or not anyone opens the notes pane. Plus the same properties as a spreadsheet.	The notes are the reason a deck can be presented a second time. They are supposed to be there — in the copy the firm keeps.
Word documents	The same authorship and revision properties, plus whatever tracked changes and comments were resolved rather than removed.	Collaboration leaves a record, and accepting a change is not the same act as deleting the record of it.
All of them	Hidden rows, hidden sheets and hidden slides. The formula behind a total. The full-resolution original of a picture that was cropped in the document rather than before it.	Hiding something is a display instruction. The content stays in the file, addressed by anyone who looks past the view.

The last row is the one that surprises people most often. Cropping a photograph inside a document usually adjusts what is drawn rather than what is stored, and hiding a column sets an attribute rather than removing the numbers. Both are presentation decisions that a recipient can reverse with no special tooling and no ill intent — a stray click on *unhide* is enough.

3. Four ways it gets out

These are the recurring patterns, in the order they tend to occur.

The name on the file

A document sent as the firm's work product carries the name of whoever happened to create it, along with the name of their machine and the software they used. In a dispute, correspondence and filings acquire an authorship trail nobody intended to publish — and in a firm that rotates work between offices, an unexpectedly informative one.

The title it had before

Renaming a file in a folder does not rename the file. The title inside it is a separate field and is left behind — so a document circulated under a neutral name can still announce, inside itself, the client or the transaction it was originally drafted for.

The workings behind the figure

A schedule sent as a summary often arrives with every formula intact, which means the recipient has the assumptions, the intermediate steps and any adjustment that was made late. Where the sheet has hidden rows, they have those too. What the sender believed they were sending was the conclusion; what arrived was the reasoning.

The note under the slide

A deck sent as a handout carries its speaker notes. Notes are written for the presenter and are, by their nature, franker than the slide — the caveat, the number that is not yet firm, the instruction not to raise something. Every one of them travels in the file, and the recipient does not need to do anything unusual to read them.

None of these requires an adversary. They require an ordinary recipient who opens the properties panel, clicks unhide, or presses the button that shows the notes.

4. Why taking it off is more than one operation

The instinct is that this is a single checkbox. It is not, and a tool that treats it as one leaves a file that is either still carrying the material or is quietly damaged.

There is usually a second copy, in a different place

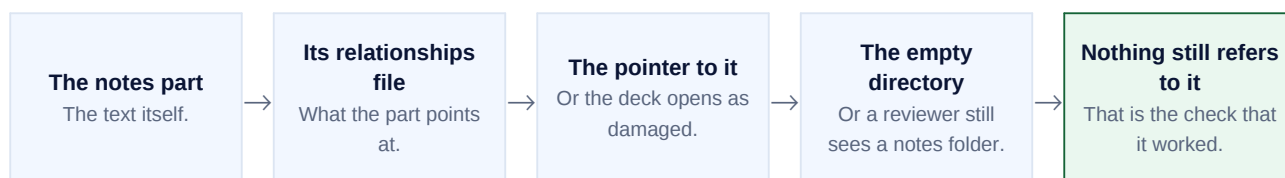
A PDF states its authorship twice: once in the information dictionary, and again in an XMP stream that keeps its own history of the same fields. Clearing one and leaving the other is the common failure, and it is invisible — the properties panel most people check reads from the first.

Oxofolio removes the information dictionary wholesale rather than selectively. That is a deliberate choice: the dictionary is not a fixed schema, and a producer is free to write keys nobody has seen before. Enumerating the fields you know about is how you leave behind the one you did not.

Removing a part means removing what points at it

Office formats are archives of parts with a manifest. Taking speaker notes out of a deck properly is four operations, not one — the notes part itself, the relationships file belonging to it, the pointer to it from the slide, and the declaration of it in the archive's own content types.

Do only the first and the file declares a part it no longer contains, which is a deck the presentation software reports as damaged. Do the first three and a reviewer listing the archive still sees a notes folder and has to be told it is empty, so the bare directory entries go too. **The check that the removal worked is that nothing in the file still refers to what was taken out.**



Four operations and a check, for one line in a firm's policy that says to remove speaker notes before a deck is shared.

The notes master is deliberately left alone. It holds the layout a note would be printed in and no note text, and removing it means editing the presentation part as well for nothing gained — a change that carries risk and delivers none.

The output is a copy, and the working paper keeps everything

These operations write a new file alongside the original, marked as the shared one. The deck the firm keeps must retain its notes, because the notes are the reason anybody can present it again, and the working paper must retain its formulas because it is the working paper. What is being produced is the version that leaves — not a cleaner version of the thing you rely on.

5. Two decisions that are not the software's to make

A cleaning pass that removes everything it can find is easy to build and produces a category of harm that is worse than the one it prevents. Two cases are reported rather than acted on.

Attached files are counted, not removed

A PDF can carry other files inside it. One of them may hold exactly the material that was just taken off the pages — and one of them may be the substantive evidence somebody deliberately attached, without which the document is incomplete. Those two cases are indistinguishable to software.

So the attachments are counted and reported, and the decision belongs to the firm. The purpose of the report is to make sure it is an informed decision rather than an unmade one.

Only the actions that execute something are taken

Embedded scripting and the actions that fire when a document is opened are removed. But a great many documents carry an open action that simply says *start at page one*, and stripping those would change how ordinary files behave for no security benefit at all. The distinction is drawn on whether the action executes.

And a third, where flattening is offered

Formulas in a spreadsheet can be replaced with their values, so a schedule goes out as figures rather than as reasoning. It is optional, and it works from the value the file has already cached — **nothing is recalculated**. A formula whose cached result is missing is left alone and counted rather than being replaced with a blank.

That last rule is the whole design of the feature. A tool that wrote an empty cell where a total used to be would produce a schedule that silently adds up to nothing, delivered to a client, looking finished. Refusing the cell and reporting it is the only safe behaviour, and it means the count is something to read rather than ignore.

6. What this does not reach

The pass covers PDFs, spreadsheets and presentations. Its edges are worth stating precisely, because a reader who assumes a wider scope is left worse off than one who never ran it.

- **Word documents are not sanitised here.** They can be inspected for personal identifiers before they are sent, but there is no metadata-stripping pass for them. Convert to PDF and clean that, or do it in the word processor's own document inspector.
- **Hidden rows, hidden sheets, tracked changes and comments are not removed by this pass.** It clears the document's properties; it does not decide that hidden content should stop being hidden. Both are real exposures and both remain the reviewer's job.
- **This is not redaction.** Taking the authorship off a document does nothing to the identifiers printed on its pages. They are different operations against different parts of the file, and doing one is not evidence of the other.
- **It cannot know what is sensitive.** Properties are removed because they are properties, not because their contents were assessed. A title that names a client is removed exactly as a blank one is.
- **The original still exists.** A copy is produced for sending, and the source keeps everything it had. That is the intended behaviour and it means the unsanitised file remains in the folder, subject to whatever governs the rest of the folder.

The useful way to read this list: the pass handles what accumulated automatically. What somebody chose to put in the document, and then chose to hide, still needs a person.

7. A pass to run before anything leaves

Independent of tooling. Six checks, most of a minute, on the copy you are about to attach rather than on the working file.

- **Open the document properties** and read every field. Author, company, title, subject, keywords, last modified by, and any custom property. Look at the title especially — it is the field that most often still names something the filename no longer does.
- **In a spreadsheet, select every row and column and unhide.** Then check the sheet tabs for hidden sheets. Do this on the copy, look, and decide.
- **Click on the totals.** If a formula appears in the bar, you are sending the workings. Decide whether you meant to.
- **In a deck, open the notes view** and read the notes on every slide, not the ones you remember writing.
- **In a document with a review history,** confirm that changes were removed rather than accepted, and that resolved comments are gone rather than hidden from the current view.
- **For a PDF, check for attachments and for anything scripted** — and remember that the properties panel typically shows you one of the two places authorship is recorded, not both.

The organisational version of this is shorter and works better: make the sanitised copy the artefact that gets attached, so the file that leaves is one produced for leaving rather than the working file with a review pass over it. A convention beats a checklist, because a checklist depends on the day somebody is having.

8. Conclusion

Every document that leaves a firm carries a description of how it was made. Most of the time that is harmless and nobody looks. The exceptions are not rare enough to be interesting, and they are never discovered by the sender.

What makes it tractable is that the material is finite and it is in known places. It can be listed, taken off, and the removal checked by confirming that nothing in the file still points at what was removed. The parts that cannot be decided by software — the attachment that might be the evidence, the total whose formula might be the point — are few, and they are better handed back than guessed at.

The page was reviewed. The question this paper asks is whether the file was, and it is answerable in a minute on the copy in front of you.

This paper describes what software does. It is not legal advice, and nothing here is a statement about your professional obligations when disclosing or filing documents.