Beratung

Optical Character Recognition – OCR

23. April 2014, von Irène Wilson

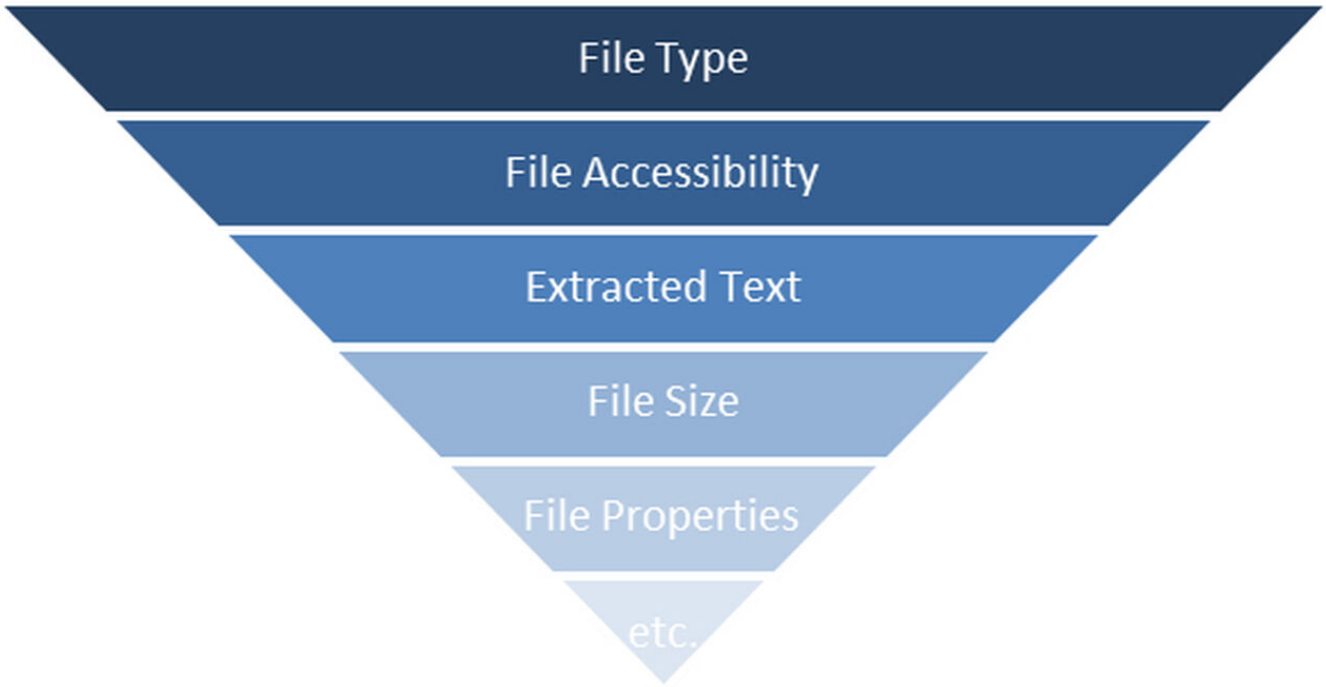
Introduction

Most forensic tools allow you to easily search through a text file or a simple document. However, an image could potentially contain text. This is typically the case for non-searchable PDFs, scanned documents, screenshots, etc. For this kind of data, Optical Character Recognition (OCR) is needed to recognize and extract the text contained in the image. Without this step, all searches you run might return incomplete results, as all these non-searchable documents will not be taken into account. Therefore, to guarantee a high level of completeness and reliability of the results, OCR is a critical step in your workflow.

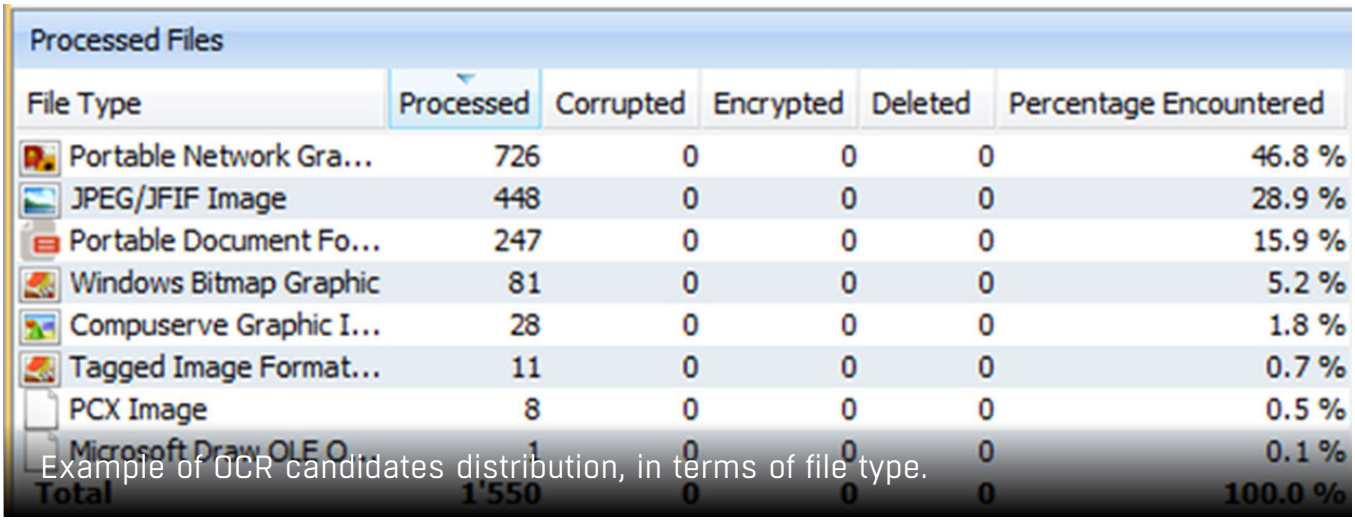
Selecting items for OCR

Even though the usefulness of OCR is no longer questioned, the selection of the documents candidates for OCR has proven a more complex matter. A large variety of criteria can be

taken into account and every expert has their own “OCR candidates” recipe. No universal solution exists, as the best OCR strategy is tailored to the project means and needs. The time and cost impact of OCR should always be put in relation to its need to fulfill the client’s expectations and requirements.



The most basic criterion is usually the file type. Images and PDF files are often considered when it comes to OCR. This selection is made even easier with tools such as Nuix, which extracts embedded images or documents. Typically, an image being part of a Word document (embedded file) would appear as a distinct item and would therefore be easy to highlight. Whether all types of images should be OCRed, just some of them or none always depends on your project. How likely is it, that information relevant to your case is contained within an image? Typically, while a fraud would shift your focus towards booking data, for a data theft you might want to take screenshots into account. You might however wonder whether all types of images are relevant. Some company scanners have specific output formats you want to focus on. If your client uses a specific screenshot software, knowing its output formats would help you to optimize your selection. A good knowledge of your client environment and policies can have a real impact on your OCR strategy. Usually, you want to present the different options to your client and explain their impact in relation to their expected relevance.



Processed Files

File Type	Processed	Corrupted	Encrypted	Deleted	Percentage Encountered
Portable Network Gra...	726	0	0	0	46.8 %
JPEG/JFIF Image	448	0	0	0	28.9 %
Portable Document Fo...	247	0	0	0	15.9 %
Windows Bitmap Graphic	81	0	0	0	5.2 %
Compuserve Graphic I...	28	0	0	0	1.8 %
Tagged Image Format...	11	0	0	0	0.7 %
PCX Image	8	0	0	0	0.5 %
Microsoft Draw OLE O...	1	0	0	0	0.1 %
Total	1'550	0	0	0	100.0 %

Example of OCR candidates distribution, in terms of file type.

Excluding encrypted and corrupted files is probably the most commonly agreed upon criterion. A downside of this could be files that were wrongly flagged as corrupted. Unfortunately, this is not a theoretical assumption but a real-life experience. In some cases, PDF files flagged as corrupted by forensic tools can actually be opened in their native format and even OCRed. Therefore a proper quality control of your data is crucial.

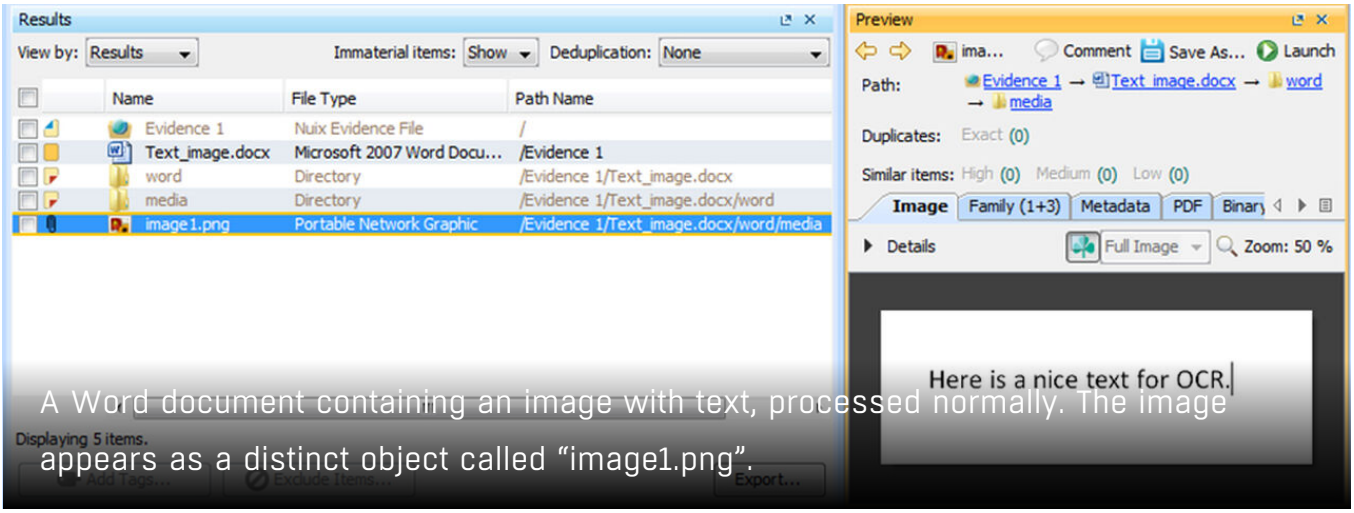
Another more ambiguous criterion is whether the candidates should be limited to the files devoid of any text. While these files are definitely good OCR candidates, this rule might be considered too strict in some circumstances. There could be PDF files where only the headers or footers were extracted. Therefore a maximum character count might be more appropriate. However, there are even worse cases where the extracted text doesn't match what you see on the screen and is illegible, because of encoding issues. This kind of files will have a high enough number of characters and slip through your filtering process. In the end, as project needs, cost and time are brought into consideration, best effort is the only correct answer.

The file size can also be used as additional filter, even though this requires detailed knowledge of your data and some time to sample it. By excluding the files which are most likely too small to contain any text, the number of candidates can sometimes be massively reduced. An upper limit is also an option depending on your case, as very good quality pictures are usually less likely to contain text.

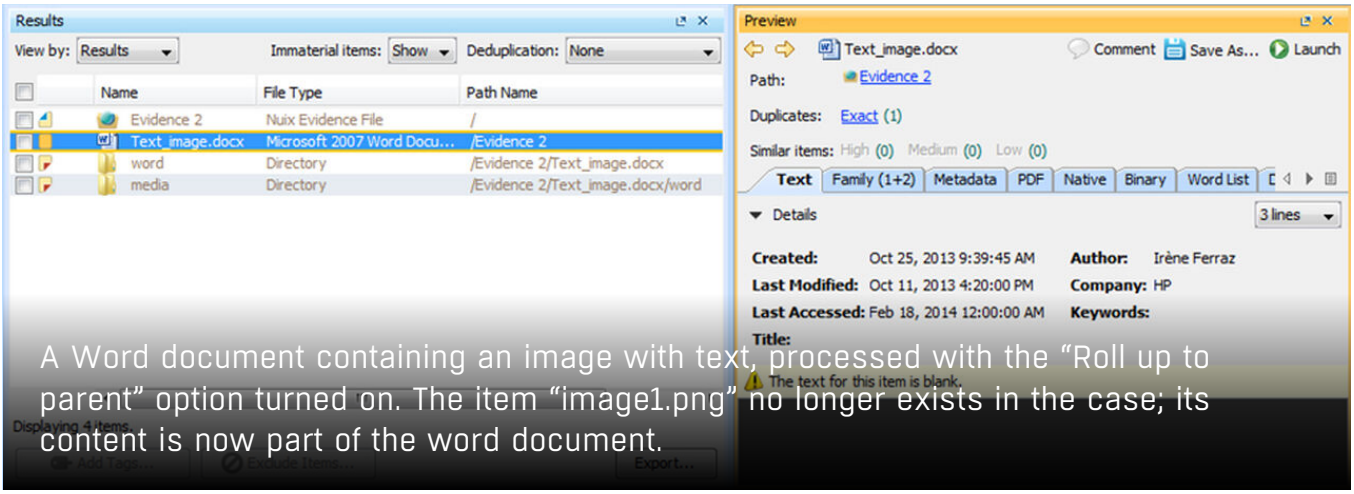
While on the topic of pictures, if images created with a camera are irrelevant to your case, you will exclude files containing EXIF or GPS properties. Although, according to my experience, this often only has a minor impact on reducing files.

For Nuix users, an additional choice appears with immaterial items. While most people tend to ignore them, it is nevertheless important to be aware of the fact that immaterial items can contain text or images. Therefore, they shouldn't necessarily be excluded from the OCR

candidates. Furthermore, a processing option in Nuix, “Hide immaterial items (text rolled up to parent)”, allows the user to hide the immaterial items without losing the text they might contain, as it is added to their parents. So what about the immaterial images? After testing, we noticed that they are rolled up to their parents. While you might not have the impression of losing information this way, it does have an impact on your OCR selection. This fact should definitely not be overlooked. How are you going to identify or even send an image to OCR, when it is attached to a Word document? Are you going to keep every Word document as potential OCR candidate? And what about searchable PDFs containing some graph or table as image? As eDiscovery is all about best effort and solutions tailored to the needs of a project, the impact of every decision should be known and understood.



A Word document containing an image with text, processed normally. The image appears as a distinct object called “image1.png”.



A Word document containing an image with text, processed with the “Roll up to parent” option turned on. The item “image1.png” no longer exists in the case; its content is now part of the word document.

OCR Limitation

Once you have selected which documents to OCR, you want to deduplicate them to reduce the necessary time to OCR. There are then a number of different ways to perform OCR. Some forensic and eDiscovery software packages have this as a built-in feature while you can also choose to use specialized OCR tools. Some image the documents before applying OCR while others work directly with the original files. All those options have their advantages and disadvantages, but in the end speed is often the main decision trigger when it comes to choosing your tool.

Whatever tool you go for, always keep in mind that no product is perfect. The quality of the result is strongly impacted by the resolution of the original file, its quality and its contrast. Typically some tools don't cope well with inverted contrast (white text on black background). Numbers or signs are sometimes incorrectly interpreted. Most tools also rely on language dictionaries to optimize the result. On the one hand, a document in a language you have not previously selected will provide poor results. On the other hand, the more languages you enable, the worse the quality gets. Finally, there are often documents which fail OCRing. You should keep in mind that such documents won't be part of your keyword search scope anyway and might need manual review if you want to go for the safest approach. The quality of your OCR results also impacts the reliability of your search results for these specific documents. As usual in forensics, the absence of proof is never the proof of absence. Not getting a hit using a keyword DOES NOT prove that the keyword is not present in your data set.

Conclusion

eDiscovery, unlike traditional sciences, doesn't respect strict universal rules. Best effort is the key. This has two main consequences. First, no decision is right or wrong. Secondly, you need to know the implications, benefits and downsides of any of your choices. A good decision is one that can be justified. The same applies to OCR. Whether you do it or not, and how you actually apply it is not what really matters. What makes the difference and makes you an expert, is knowing the limitations and implications of your choices. However, you can do even better. Helping the client make an educated decision on the approach to undertake will prove that you don't only master the technical component, but are also able to communicate it clearly to non-technical parties involved. Even though our field might look a bit tough to non-forensic people, presenting it as a complete black box is not helping your

client. Implementing a project in a collaborative manner, getting input from your client and involving them in the decisions will ensure that you correctly understood the client's expectations and achieve high quality work.

Irène Wilson

Irene Wilson ist auf digitale Forensik und eDiscovery spezialisiert und hat im Laufe ihrer langjährigen Erfahrung für Kunden aus einer Vielzahl von Branchen in ganz Europa gearbeitet. Zu ihren Qualifikationen gehören die renommierten Master-Titel für Nuix Workstation und Nuix Discover.

Swiss FTS AG | www.swiss-fts.com | +41 43 266 78 50 | info@swiss-fts.com

https://swiss-fts.com/blog/optical-character-recognition-ocr