Technisches Wissen

Advanced Deduplication, Twins beyond Fingerprints

05. Mai 2020, von Irène Wilson

Introduction

Technology is used to support investigations and review projects in many ways, including saving time and money through data reduction. While basic duplicate removal through hash value comparison is well accepted, it has limitations that powerful tools and creative workflows can overcome. Here is an overview of some of those techniques.

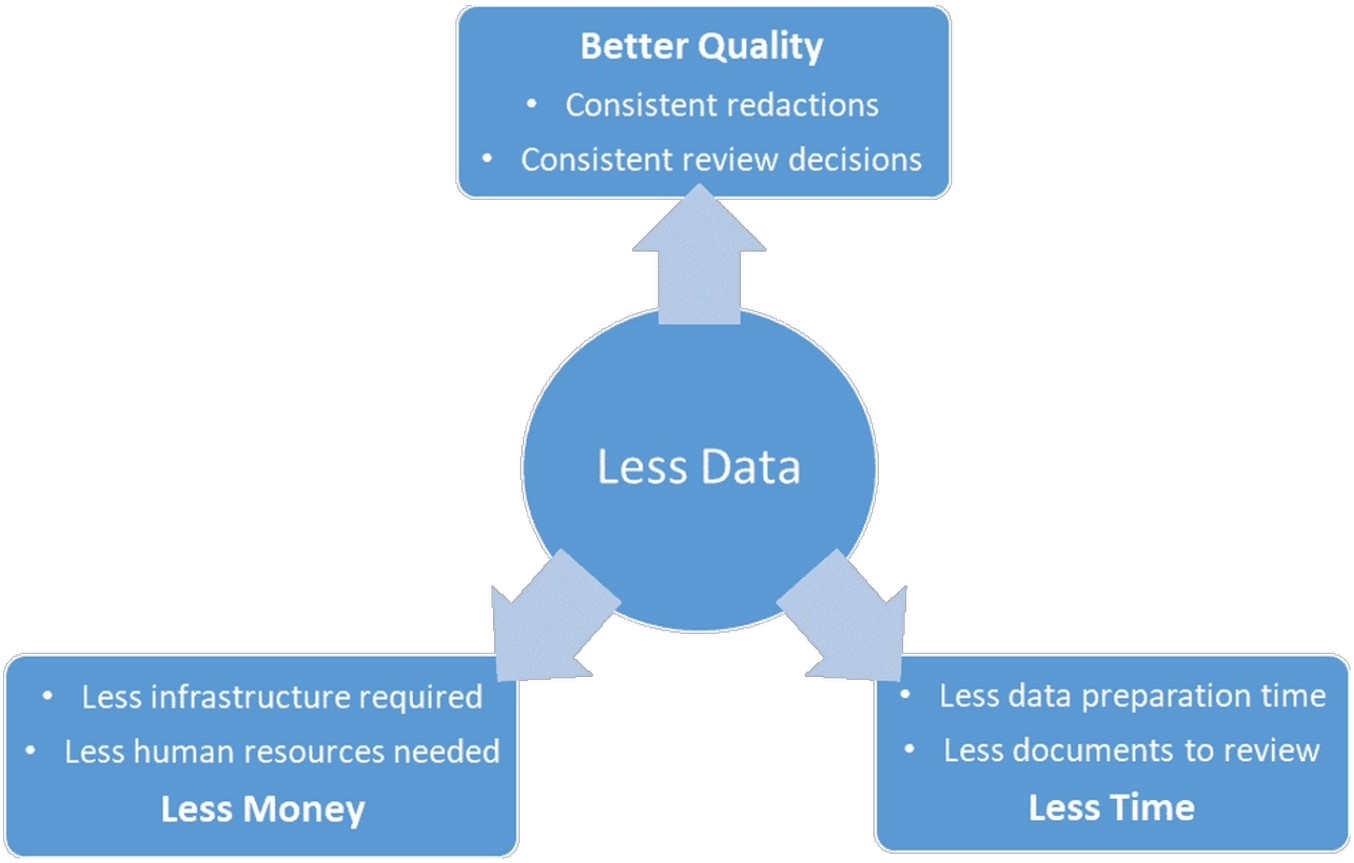
Data reduction relevance

Before debating the pros and cons of the various data reduction approaches, let's first consider why we are doing this. A larger volume of data requires more storage, more powerful hardware, more data preparation time, and more review time, with the risk of missing relevant information because it is obscured by background noise. Time is money, and therefore more data has a double impact for costs in terms of hardware and both personnel

and machine processing time.

This doesn't even take into account other considerations such as deadlines. What's the impact of missing a deadline? A big fine most likely, with a lot of zeroes. Data reduction helps you reduce time and costs in your investigation.

There is also a positive effect on quality. Reviewers are human and therefore make subjective decisions. Having duplicates in your data set often results in inconsistent review decisions, where duplicates of the same file can be coded differently depending on who performed the review. This gets even trickier when redactions and pseudonyms are involved. Data reduction will improve review consistency and quality.



Weaknesses of traditional hash value deduplication

It is common to use hash values to remove duplicates in eDiscovery projects. Those hash values are computed by applying an algorithm to the binary content of the files. MD5 hash values are a common output of such process, and are then used to identify and exclude duplicates. Such values are often referred to as "digital fingerprints".

Did you know that identical twins have different fingerprints? In a similar manner, files that

from a reviewer's point of view are identical can have different hash values. Two factors explain those differences: information that is not visible to the user, such as properties or technical information stored within the file but hidden from non-technical users, and the formatting of the information displayed to the user, such as compression or text formatting. Because of those weaknesses, a simple hash value deduplication often leaves a bitter after taste to the reviewer, who feels like they are performing redundant work and that the deduplication process hasn't worked.

Common commercial solutions

"Purified content" hash values

Several eDiscovery software vendors have developed their own solutions to provide effective deduplication when it comes to emails. Electronic messages are extremely tricky because of their header, which is impacted by the route that the message takes from sender to receiver, and the formatting that software can apply to their body. To tackle this challenge, tools such as Nuix strip the headers of irrelevant information, and clean up the body before applying a hashing algorithm. This allows a message taken from the sender's mailbox and the same message in the recipient's mailbox to be identified as duplicates. Although this solution is reproducible, it remains proprietary and therefore doesn't allow comparison across different tools.

Text-based deduplication

The solution described above is specific to emails, but it doesn't overcome the different formats information can take. An email for example could also exist as a PDF or an image. To catch those, the best approach is to ignore the shell and focus on the pearl: the information itself. Computing an MD5 of the unformatted text extracted from the file allows for the comparison of information stored in different formats. Shingles and near-duplicates bring this concept even further allowing for a chosen degree of similarity in the text, instead of an exact match. This can help get around OCR inaccuracy, but it will also group documents based on the same model (such as a template email with only some values being updated before sending), or different versions of a same document. It is important to clarify the purpose of deduplication, and to verify that this method gives the required results. Removing documents based on a same model because your tool and approach flags them as duplicates could do a lot of damage depending on the case.

Image solutions

When it comes to images, there is a well-known issue to police officers: MD5 doesn't allow for effective deduplication, because every image appears multiple times on a system, in different sizes, formats and levels of compression. PhotoDNA is a very clever algorithm to deduplicate images, which is now implemented by most tools on the market. It can identify duplicate images while accounting for these issues. However its use is limited to qualified organizations and law enforcement agencies.

Binary fuzziness

Another file type creating issues is malware, where the code is obfuscated, or slightly changed, so that the binary data appears different even though the functionality remains the same. The algorithm SSDeep assesses similarity on the binary level to bypass this limitation in identifying duplicates.

Additional real-life email issues

All the technologies and algorithms mentioned above do a great deal to humanize the deduplication process, identifying duplicates in a way that a reviewer or investigator understands them rather than how a machine would identify them, with a more flexible and pragmatic approach. Even with these improvements, when it comes to document review this is still not good enough. Experience has shown emails to present more challenges than any other file type.

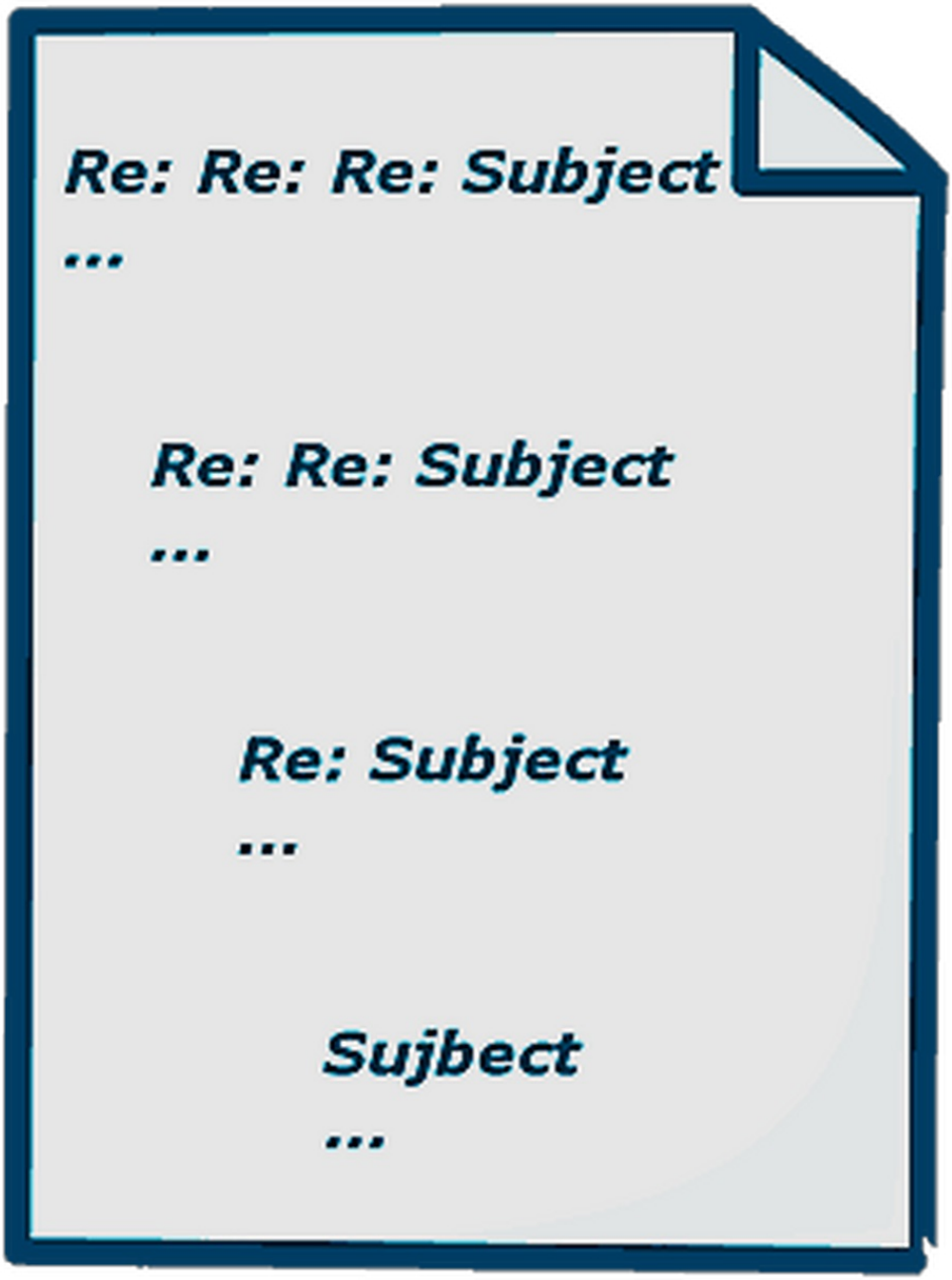
In addition to the multiple paths that an email can take, different systems store emails in ways that can impact their hashes. Emails are sometimes stored in databases, where it is common practice to strip the attachment so that they can be deduplicated to save storage space. When collecting the same email from different sources, forensic and eDiscovery tools do their best to rebuild the message as close as possible to the original, but the road is full of obstacles that fool even "purified-content" hash values. Email addresses viewed from an external versus internal viewpoint, attachment order, and archived message warnings are all slight differences that current tools can't handle by themselves. That's when the forensic investigator can demonstrate the depth of their art by customizing workflows, developing tools, and creating new approaches to tackle, in a reproducible and controlled manner, the new challenges found on this unpaved road. The result feels miraculous, particularly when data was collected from all possible sources to cover any potential gaps, and the resulting

level of redundancy is very high. The ideal approach will also take into account the quality of the various sources available, and give priority to the best one when given the choice.

Email threading

Pushing the concept of “duplicate” further than most technically-minded individuals would feel comfortable with, reviewers often complain about looking at several messages from the same conversation thread. Most tools on the market now allow email thread analysis. The concept is simple enough: keep the most inclusive email in the chain, as well as any message with additional content (including attachments). The proper implementation is a bit trickier and can have severe pitfalls. It is therefore important to know your tool well and understand the fine details of the technical process.

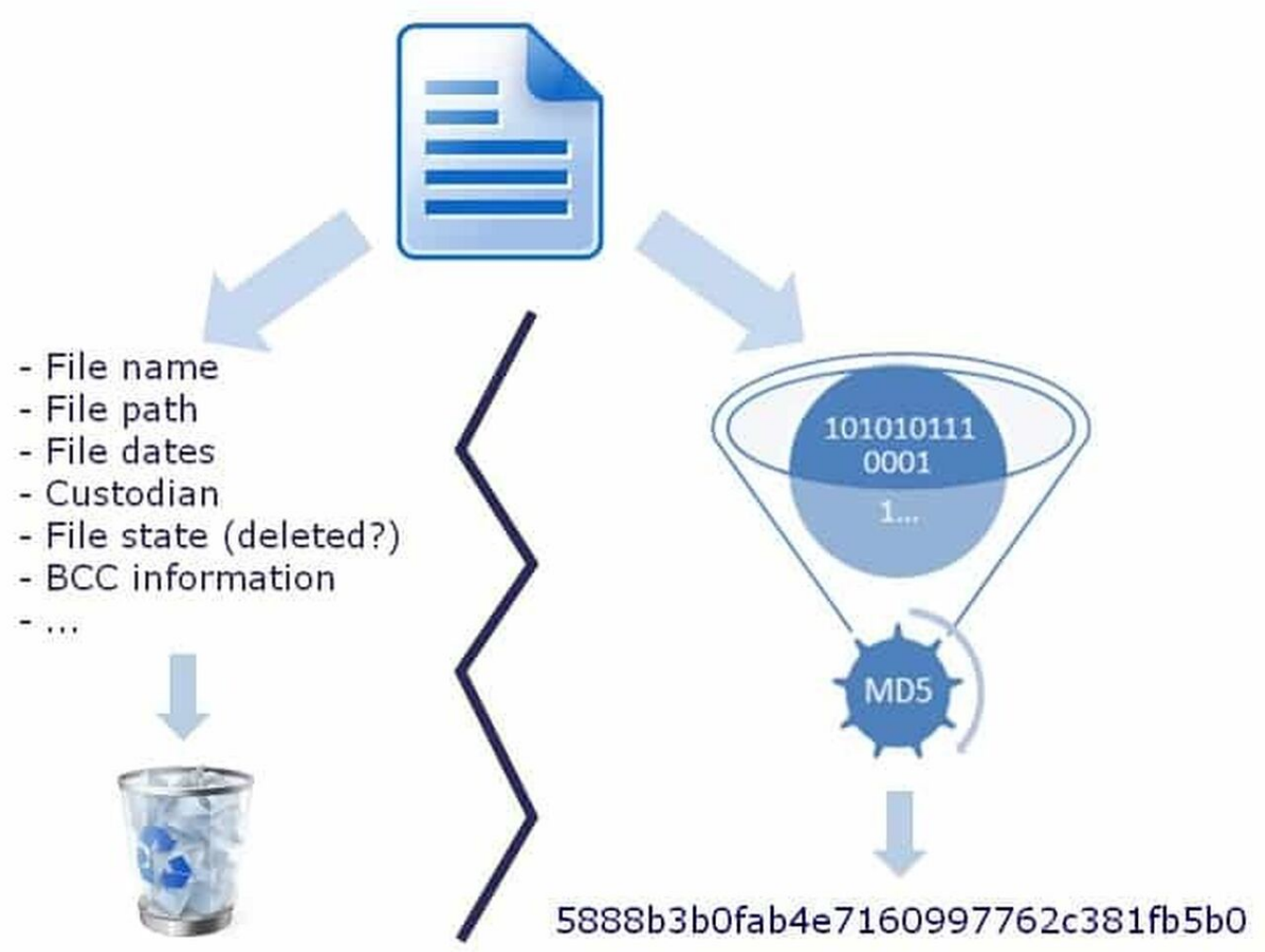
Pushing the concept of “duplicate” further than most technically-minded individuals would feel comfortable with, reviewers often complain about looking at several messages from the same conversation thread. Most tools on the market now allow email thread analysis. The concept is simple enough: keep the most inclusive email in the chain, as well as any message with additional content (including attachments). The proper implementation is a bit trickier and can have severe pitfalls. It is therefore important to know your tool well and understand the fine details of the technical process.



Data reduction risks

Data reduction is necessary, there is no doubt about this. However, it can have negative consequences and therefore needs to be planned carefully and implemented correctly. The specifics of the investigation, for example whether contextual information such as a file path impacts the pertinence of a document, needs to be shared with the technical team, so that they can adjust the approach and prevent falling into deduplication pitfalls. Even if we limit our approach to standard deduplication, keep in mind that contextual information like document families is ignored in the calculation of the hash-value, but it could have relevance in the investigation. Our previous article, "Deduplication Hidden Downsides", sheds some

light to the shadows of such a process.



Here are some practical examples which highlight how real the risks are of not taking context and other pitfalls into consideration:

In a data leakage case, the reduced data set was searched to find the leaked information. The file in question was found on a share drive, in an acceptable location based on the client workflow for that type of information. When going back to the full data set, copies of that same file appeared to be located in a user folder, pointing directly to a potential suspect.

When realizing the size of the reduced data set, even after limiting the documents to review based on keywords, the client decides to apply date filtering. The case containing loose files, part of the data relevant for the chosen period would be overlooked if the date filtering wasn't applied on the complete data set first.

An investigator found illegal content in a recovered picture from the unallocated area of a suspect's hard drive. Evidence found in such location is tricky to present to court, as it is

lacking context and it's impossible to deduct any user intent to download, save or copy the file. However, if a duplicate of that same file is available in the Pictures or Downloads directory of the user profile, then the impact of the evidence is completely different.

False positives are another risk in duplicate identification. While this risk is extremely low with a standard deduplication approach, it gets higher with more advanced features. As mentioned above, text-based deduplication allowing some flexibility could identify different versions of the same form as duplicates.

The "purified content" hash value also has its pitfalls. As an example, Nuix applies that specific approach on calendar items, however it is common to have several calendar entries with the same author, recipients, subject and bodies, but occurring on different dates. Who has never registered recurring doctor appointments in their agenda this way? Considering those different entries as duplicates and removing them could end up in overlooking a possible alibi.

Conclusion

Deduplication is well accepted and necessary nowadays, although the simpler methods are often insufficient and professionals have been applying more advanced approaches for some time. Having the client look at you as if you have just saved the day because you could reduce the amount of data by 70% is always a nice feeling, but if this comes with the risk of never finding the smoking gun, the gratitude won't last. Communication between the investigators and the technical staff is key to a safe implementation of such technics.

This article focused on data reduction, but data prioritization is becoming the next stage in this battle between investigators and increasing data volumes. Whether it is through algorithms extending reviewer decisions to a larger data set, training a system to accurately assess the remaining data, or reassessing the relevance of the remaining data on the fly as the review progresses, software developers and eDiscovery experts keep coming up with innovative approaches.

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.