Beratung

Nuix and the MD5 Conundrum

23. Oktober 2014, von Irène Wilson

Introduction

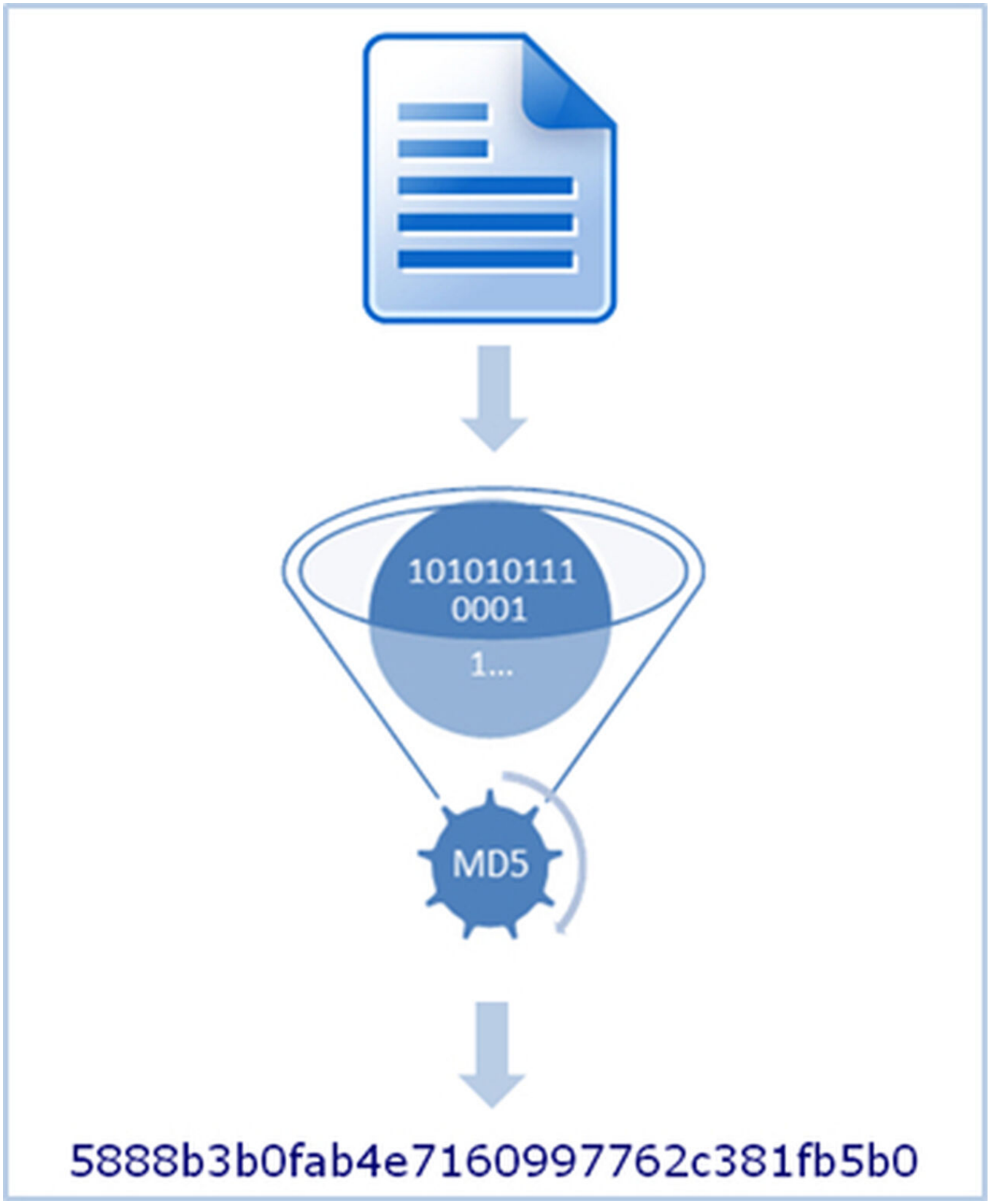
MD5 hashes are usually described as “electronic fingerprints”. They are widely used as unique identifiers for electronic data. They allow to testify the integrity of files or to prove that two files have identical content. While the original frame of MD5 digest is really strict and reliable, the practice can sometimes be less rigorous, as will be presented below.

Even though MD5 is explicitly mentioned, the topic covered by this article is broader and can be extended to other hashing algorithms. Therefore, the risk of collision will not be treated here.

Definition

An MD5 digest is a 128-bit value which is usually presented as 32-digit hexadecimal number. It is computed by hashing the binary content of a file with the MD5 algorithm. The algorithm

is such that a single bit difference will result in a very different digest and that it is not possible to guess the content of a file based only on its digest. Furthermore, as it only depends on the content of the file, it is therefore completely independent of its context, e.g. file name, path or dates. Thus, the resulting digest is completely portable and reproducible.



Limitations

In practice, the MD5 is widely used to prove the non-alteration of a file or the similarity of two files. While this sounds pretty intuitive, it actually suffers from several limitations. One problem well known by police investigators is the impact of formatting or compression on videos and pictures. While MD5 digests are a great way to point out relevant material in pornography cases, any change of format or compression changes the MD5 hash, disabling the automatic identification of the file. Another issue impacting more the eDiscovery side is the diversity of email files. With emails, the header varies depending on the servers the email goes through. The client application used to receive and read emails can also have an impact on the format it applies to them. Also, the BCC field does not appear to all recipients. Therefore, the strict definition of MD5 digests does not allow recognizing the email in the sender's and in the recipients' mailboxes as being identical. Strictly speaking, they are actually not identical. However, from a logical point of view the differences are negligible. As a consequence, some tools compute "custom" MD5 digests for emails, restoring the identification capability but losing the portability as a consequence.

Nuix Particularities

Nuix is one of the tools addressing the email issue with a custom MD5 solution. But this is not the only specificity of this tool when it comes to MD5.

Loose Files

No surprise with loose files; the hashing practice follows agreed upon conventions in this case. The whole file binary is hashed to compute an MD5 digest that is completely portable and reproducible. Any other tool on the market would come up with the same result.

Emails

Regarding emails however, Nuix MD5 hashes are proprietary. The algorithm itself does not change, but the file is adapted beforehand. The documentation for Nuix (5.0.5) contains the following details in this regard:

"Since not all email types actually have a binary stream and two copies of the same message

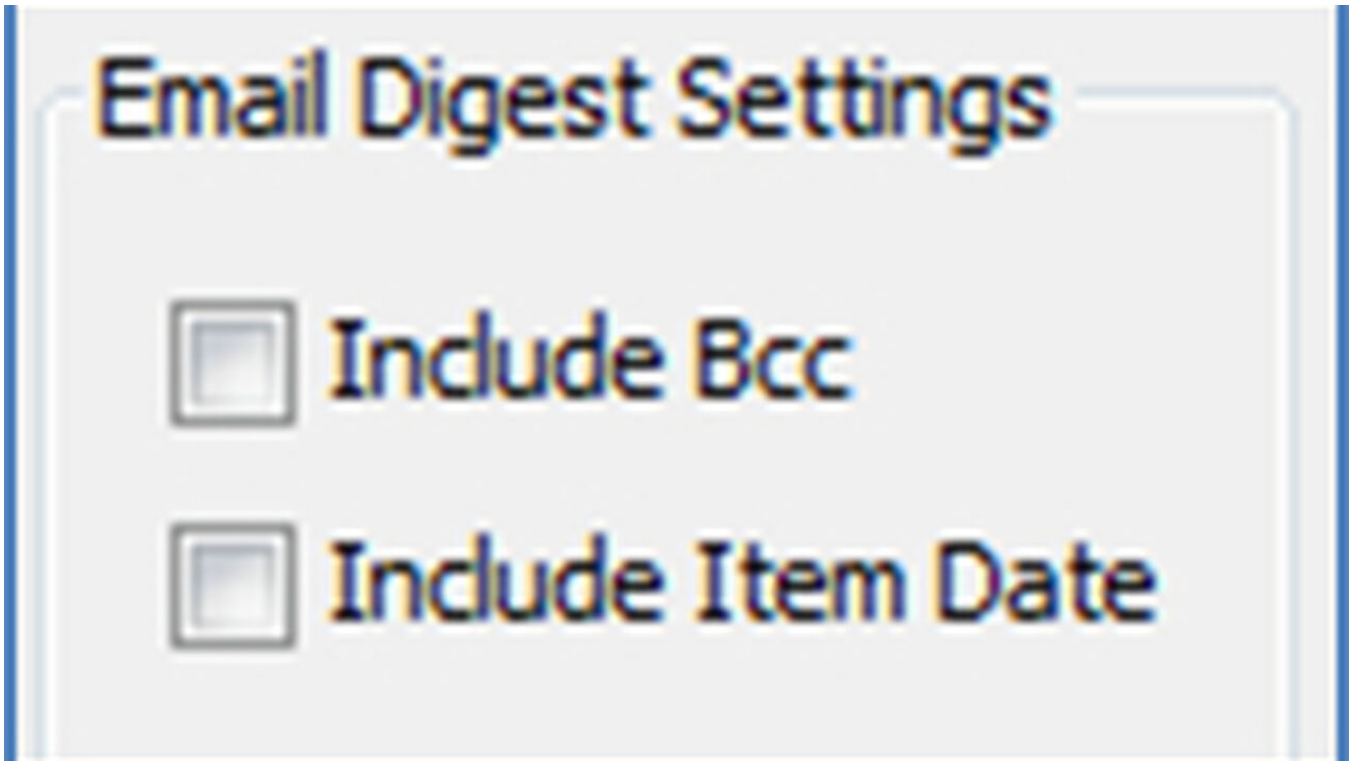
can have completely different header information, we compute an email's MD5 digest by taking the following data encoded using UTF-8 as input:

1. Subject header
2. From header
3. To header
4. Cc header
5. Email body text tokenized so whitespace and irrelevant characters are removed
6. Binary streams of all attachments

For address headers the personal part is discarded an only the address part is used. The email body is tokenized to ignore white-space differences, which can be a factor when comparing HTML and plain text messages."

Since version 4, the user can choose upon processing to include as well the BCC field and/or the item date (sent or received date) in the MD5 calculation.

Although this is useful for deduplication, it has the downside to disable the option of reproducing the digest with any other tool.



Items without MD5

Obviously, items without content like directories do not have an MD5. There is indeed no binary stream to hash. You might expect this to be linked to the immaterial flag (please refer to our article "Nuix Immaterial Items, the Grey Area" for more details about the immaterial concept). This is however completely unrelated, as some of them (mostly embedded objects) actually have an MD5 hash. Even though they are not independent of their parents, they still have content, and therefore a binary stream to hash.

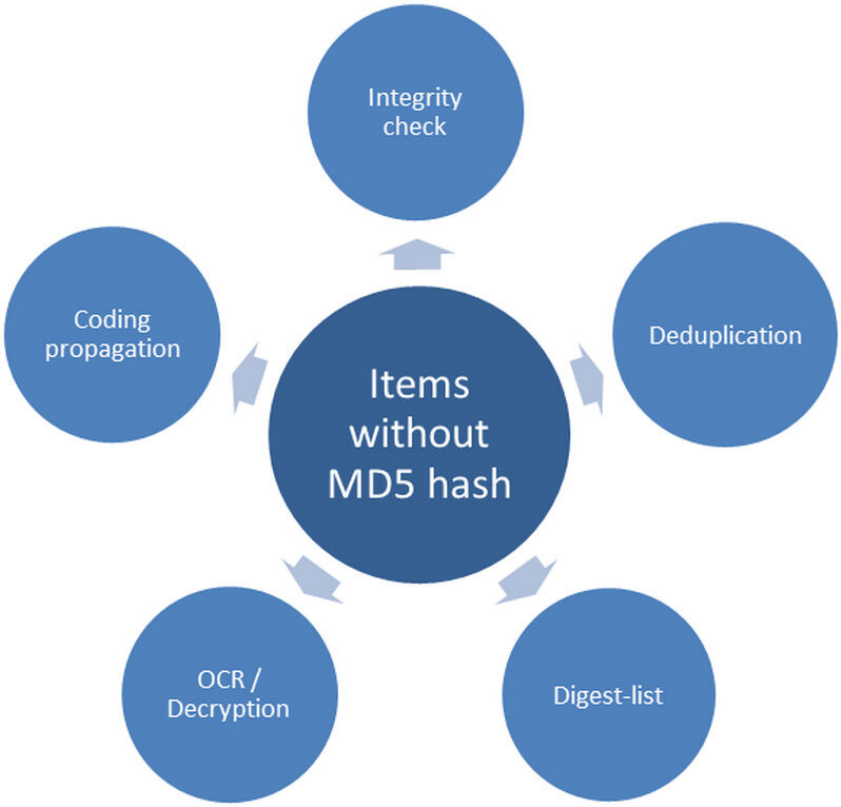
Regarding the material items without an MD5 digest, the first cause to investigate is their size. Have you ever noticed the "maximum digest size" setting when processing? This value, set by default to 256 MB, restricts the items hashed according to their size. Typically, if you keep the default value and your case contains videos, it is quite unlikely that they will have an MD5 hash.

However, there are more material items without MD5 hashes. Some of the impacted files are actually of little interest: "Microsoft Outlook Property Block", "Inaccessible Content", "Empty File", etc. While "Inaccessible Content" or "Empty File" items have reasons to be lacking MD5 digests, "Microsoft Outlook Property Block" items actually have a binary content (even extracted text). Therefore it's only a choice of implementation to not calculate an MD5 hash for them. Anyway, this probably does not have a big impact on your analysis.

More recent supported file types generate unusual children: Internet or browser history and Skype data. If you have not tried processing this type of data with Nuix 4 or above yet, be ready to lose your marks! Those types of files have a database-like structure. Nuix extracts their content and creates a new item for each entry. While this behavior can be useful for analysis, it is highly surprising to notice that those items do not have an MD5 digest. They do have binary content, but a choice of implementation made them digest-less. Unlike the material items mentioned above, these are more likely to be used for further analysis or review. The lack of a digest can therefore have a real impact.

Impact of Items without MD5

Knowing about a fact is one thing. Understanding all its implications is a totally different matter. Even though some are aware of the possibility of missing MD5 digests, unfortunately only few have a comprehensive understanding of all consequences of this situation. So here is a quick list to open your mind on some software limitations induced by the lack of MD5 hashes:



1. **File integrity:** Lacking an MD5 hash obviously does not modify your file but what if someone questions the file integrity? How are you going to prove it? Courts might also request produced documents to have a digest to allow integrity checks.
2. **Deduplication:** Files without MD5 cannot be deduplicated. This is especially annoying when your dataset contains movies that are too big to have an MD5 hash and therefore will not deduplicate. While deduplication allows you to reduce your data volume drastically, it fails on the biggest items in this case.
3. **Coding propagation:** Propagation is often used in eDiscovery to ensure consistent coding. This way, a decision made on a file will be automatically applied to all its duplicates. Unfortunately, with items lacking an MD5 digest, no propagation is possible and the coding consistency QC will not be possible, neither.
4. **Digest-lists:** Digest-lists can be used for QC purposes, deduplication of a later data delivery, or identification of specific files within another dataset. However, an item without an MD5 hash will not be part of a digest-list. In terms of QC, this is a big limitation. When reprocessing a case, you might want to use a digest-list to check that the initial case was not missing any item. Lacking MD5 hashes, you will not be able to run that check. Deduplication, as mentioned earlier is impossible without an MD5 hash as well. You will run into the same situation if you wish to identify known files within another dataset based on MD5 hashes.
5. **Decryption/OCR:** To improve performance for decryption and OCR, it is current practice

to deduplicate the files beforehand and match afterwards the results to all files sharing the same MD5 digest. This process just cannot work with items missing an MD5 hash. Furthermore, if you use a script that automatically exports selected items for decryption or OCR and rename them according to their MD5 hash, you will run into trouble.

Conclusion

In the end, even though MD5 is a well-known and recognized standard, its implementation experiences some variants. Regarding Nuix, custom email MD5 digests and items without MD5 hashes are specificities you should be aware of. We often hear that “knowledge is the key”. However, I would go further. Even though deep understanding of your tool is the key, it is apprehension of the implications of your tool’s specificities that will really enable you to provide your client with the best advice. The path to mastering a tool is long, but it is still the most exciting challenge for forensic experts.

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