PAEB_118

Plagiarism from a Digital Forensics perspective

Clare S. Johnson, University of South Wales*
Dr Ross Davies , University of South Wales

Keywords: *Cheating, Copying, Detection, Forensics, OOXML, Plagiarism*

Plagiarism and contract cheating are serious academic issues that 'undermine the integrity of education' (Bretag, 2013). There are a number of tools that can help assessors detect plagiarism – particularly where text has been copied and pasted: Turnitin (https://www.turnitin.com), PlagScan (https://www.plagscan.com) and Urkund (https://www.urkund.com) are examples of such tools. The providers of these tools are also developing authorship tools that use stylometrics and linguistics to determine matches between authors (whether the submitting author, or a third party). It is also possible for an assessor to copy passages of text and paste them into a Google search (or similar) with quotes surrounding the passage to see if there are any immediate online matches.

In a previous paper "Using digital forensic techniques to identify contract cheating: A case study" (Johnson & Davies, 2019), the authors began using digital forensic techniques to see if it was possible to detect contract cheating. In that paper, consideration was given to how a student would typically assemble an original submission, through multiple edits and rewrites, and how forensic techniques allow review of those edits through examining the Open Office extensible markup language (OOXML) format of the document and thus identify flags for unusual behaviour.

Other research into the application of digital forensics using the OOXML approach include the use of digital forensic techniques for identifying copyright issues (Fu, Sun, Liu & Li, 2011) and intellectual property (Jeong & Lee, 2017), though the latter requires access to the entire file system of the computer where the document was created, and Xiang, Sun, Liao & Wang (2016) who discuss the use of extensible markup language for transmission of secret information. (Jeong & Lee, 2017). As (Didriksen, 2014) notes: "it is desirable to connect the actions performed, e.g. editing the document, to a specific physical person or several people" when carrying out a digital forensics investigation, as this permits investigators to attribute certain actions to specific users. Thus the authors of this paper wonder if these features might be useful in attributing ownership of a student submission.

Research aim and objectives

This object of this paper is to explore the use of OOXML to see if there are other flags or features that might raise suspicion that a piece of work has not been created in an authentic way and hence may be plagiarised or contracted. Specifically, the research aims to:

1. Describe various stylistics features of OOXML;

2. Analyse which features of OOXML may be useful in determining the authenticity of a document;
3. Determine the extent to which forensic analysis of these features can help determine originality.

Methodology

Two documents were created. The first document ('originalwork') was created by opening a new blank Word document and typing in two original paragraphs of text and making some minor edits. The document was saved, one word was highlighted in bold, and then was resaved. The second document ('plagiarisedwork') was created by opening a new blank Word document, copying text from a Wikipedia page and pasting it into the document, followed by some minor formatting changes (removing bold and hyperlinks). This allowed the authors to compare the two pieces of work. Copying and pasting text from the Internet is sometimes known as 'patchworking', where students take passages of text from the Internet and build them into a submission, without giving adequate credit (Kumar, P. M., Priya, N. S., Musalaiah, S., & Nagasree, 2014).

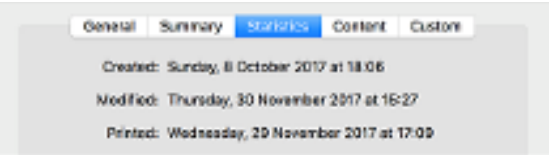
OOXML format

As discussed in the previous paper by Johnson & Davies (2019), Microsoft Word uses 'Office Open XML Format' (OOXML), where a document is created from a combination of other underlying documents. Much like a film is made up of many scenes, then has music special effects and credits added, and is finally packaged up into a single item, a Word document (docx) is made up of a number of other files, compressed into a single package.

The Open XML format has been around since Microsoft Office 2007 and was designed to bring a number of benefits to individuals, organisations and developers. These benefits include improved damage recovery because the separate components of each document are stored separately, meaning that if one component is damaged it may still be possible to open the file; better privacy and control over personal information, because sensitive information can be more easily identified and thus removed if required, and more compact: it is this feature that we can take advantage of when carrying out a digital analysis of the file (Microsoft, 2019).

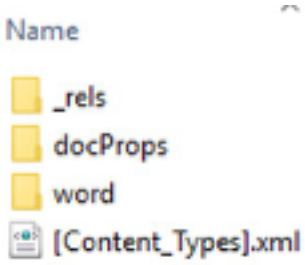
It is simple to review the document properties of a file when opened. This information can be found under the **File** menu by selecting **Properties** and then the **Statistics** tab (Fig. 1) (depending on Word version), although this data is not always reliable. Instead, looking inside the packaged contents of this document can reveal much more interesting data about the file and the way it was created.

Figure 1: Document Properties panel



To look inside the compressed docx file (or package), it first needs to be decompressed, by changing the extension of the file from .docx to .zip, and then choosing Extract, or Unzip (depending on your system). Opening the folder that is created then reveals a series of subfolders: rels; docProps; word and a single file [ContentTypes].xml.

Figure 2: Decompressed docx file



The **[Content_Types].xml** at the root of the folder contains a list of the content types of the parts within the package. The **_rels** folder tells Word how the parts relate to each other and to resources outside of the package.

Within the **word** folder, we find the following content as a minimum: _rels, theme (folders); document.xml; settings.xml; styles.xml. The file containing most of the content is the **document.xml** file, and this is the file focused on for this paper. The **document.xml** file is the main xml file for the document and includes the document's content and run identifiers. These run identifiers (RsiD tags) indicate how the document was built by placing each and every edit inside a tag, or 'run'. In the previous paper by the authors, a detailed review of the run identifies was carried out.

Flags for plagiarised work – Discussion and Results

Reviewing both revealed some interesting forensic artefacts which were considered worthy of investigation. Inspection of the document.xml file raised a number of suspicious flags.

Code inside this xml file tells Word how to render the document when displaying it on screen. Styling is defined within a <w:> namespace, which is developed by adding a relevant element. For example <w:document> tells Word that it is looking at a Word document; whilst <w:body> represents the body text of the document. A genuinely created Word document

yields a number of typical xml instructions, but in the samples some anomalies were detected. For example, the <w:cs> element refers to complex script, i.e. information about the font being used. In original work, typically a maximum of one font attribute appears within the <w:cs> tag e.g. <w:rFonts w:eastAsia=”Times New Roman”>, unless the passage where it appears is part of a field entry (e.g. Table of Contents entry) or the font has been changed (i.e. not default). However, the ‘plagairisedwork’ file includes multiple font attributes: <w:rFonts w:ascii=”Arial” w:hAnsi=”Arial” w:cs=”Arial”/>,. In addition, there are multiple font elements which may be unusual for original work e.g. <w:sz> (which relates to font size) and <w:shd> which relates to background shading. The <w:shd> tag denotes that a shadow (background) has been applied to the run. In ‘plagiarisedwork’, there are runs where the shadow is set to white, indicating that this section of text previously had text shading of some kind that needed removing. Text created within Word itself would already have a white or null background shadow, so the appearance of this command would suggest that the text had come from elsewhere.

Figure 3: Extract of xml showing <w:shd> tag
<w:shd w:val=”clear” w:color=”auto” w:fill=”FFFFFF”/>

Other tags of interest

Other tags and elements within the various documents reviewed yield similarly interesting results. The appearance of <w:NormalWeb> in some examples suggests text copied from the Internet, as does the inclusion of <w:webHidden> as this does not appear in an originally created document.

Limitations

As with many digital forensics techniques, these flag can only act as indicators. There may be genuine reasons why a document includes such flags – perhaps a student has used an online grammar tool to check their work, and downloaded an amended version for submission. Or perhaps they emailed the work to themselves. Information may have been copied from the Internet, but referenced properly, in which case analysing the flags in conjunction with the text itself is vital. These methods are also dependent on the student submitting the assignment as a word.docx file, and not a PDF or other format.

Conclusion

It is often possible to identify plagiarism through the use of text matching software, or by using search engines to find suspicious paragraph of text. However, by changing a single word, or by patchworking, students are able to outsmart tools like Turnitin and PlagScan, and to render any online searches unsuccessful. Reviewing the xml of the submission does not enable an assessor to categorically state whether work is plagiarised, but it is another option

in the toolkit for highlight ingflags which may be indicators of plagiarised work. The authors believe that there is much more that can be done in this area, perhaps developing tools which review the xml format in greater detail.

References

Fu, Z., Sun, X., Liu, Y., & Li, B. (2011). Forensic investigation of OOXML format documents. *Digital Investigation*, 8(1), 44-55. doi:10.1016/j.diin.2011.04.001

Jeong, D., & Lee, S. (2017). Study on the tracking revision history of MS Word files for forensic investigation. *Digital Investigation*, 23, 3-10. doi:10.1016/j.diin.2017.08.003

Johnson, C. & Davies. R. (in press). Using digital forensic techniques to identify contract cheating: A case study, *Journal of Academic Ethics*, Accepted for publication

Microsoft (2019), Open XML Formats and file name extension. Retrieved from: https://support.office.com/en-gb/article/open-xml-formats-and-file-name-extensions-5200d93c-3449-4380-8e11-31ef14555b18, accessed 30/11/19

Xiang, L., Sun, C., Liao, N., & Wang, W. (2016). A Characteristic-Preserving Steganographic Method based on Revision Identifiers. *International Journal of Multimedia and Ubiquitous Engineering*, 11(9), 29-38.