

Discovering unauthorised outsourcing of coursework: What DOCX files reveal

Oliver Trevisiol

06/07/23

Outsourcing to third parties, such as contract cheating providers or undisclosed text production by artificial intelligence, are considered difficult to detect. Text files in DOCX or ODT format can be used to specifically search for indications of unauthorised outsourcing. Of course, it is possible for people with the appropriate technical knowledge to change or delete information in the places presented here. In very many cases, however, this possibility is not used. On the other hand, it is also possible and perfectly legitimate to create a new and clean file for the submission version of a text. A single evidence of potential outsourcing is not yet proof, but several can give an illuminating overall picture. Further evidence can be provided, for example, by the information in the properties of electronically submitted documents.

Note for further use:



This work is licensed under a Creative Commons Attribution 4.0 International Licence.

You are free to share, copy and redistribute the material in any medium or format. You are free to adapt, remix, transform, and build upon the material for any purpose. You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

Additional information about CC licensing:










<https://creativecommons.org/licenses/by/4.0>

Discovering unauthorised outsourcing of coursework: What DOCX files reveal

Opening DOCX with ZIP

DOCX or ODT files are actually zip-archived container files containing XML files. Such archive files can be opened and unpacked with packing programmes such as 7-Zip. Often this function is offered directly in the Windows file manager after a click with the right mouse button on the DOCX file. The XML files contained therein are not particularly reader-friendly for the human eye, but can be displayed with any text editor. They contain additional information that can provide evidence of unauthorised outsourcing.

Example of the content of the Word subfolder of an unzipped DOCX file:

Name	Größe	Gepackte Größe	Geändert am
 embeddings	89 889	89 889	
 media	4 287 479	362 633	
 theme	6 793	1 571	
 _rels	1 244	316	
 document.xml	19 999	2 729	1980-01-01 00:00
 fontTable.xml	1 340	441	1980-01-01 00:00
 settings.xml	2 822	1 037	1980-01-01 00:00
 styles.xml	28 769	2 887	1980-01-01 00:00
 webSettings.xml	576	280	1980-01-01 00:00

Embeddings

The folder *embeddings* exists if the Word document contains embedded files such as Excel tables. These in turn can be opened so that information can be taken from the metadata of the XLSX file via properties (cf. also the information sheet on metadata in files).

Media

In the *media* folder you will find pictures and graphics etc. contained in the Word document. The files can also be opened and the metadata examined. Sometimes there are image files in a not yet cropped state that provide further information such as links to finding places or storage locations.

settings.xml

Word assigns a session ID (w:rsid) to all edits. A session is the editing period between two saving processes. All edits within a session receive the same session ID. The IDs are assigned randomly and in ascending order. The first editing session of a document can be recognised as w:rsidRoot. All session IDs are listed in the *settings.xml* file. If two documents have a sequence of session IDs in common, then it can be concluded that the documents have a common origin.

document.xml

The *document.xml* file contains the actual text of a DOCX file. In this file, all edits are marked with their respective session ID. If a long text contains only a few session IDs, this is a sign that the text has been worked on very little and for a short time, e.g. because finished text was copied and pasted.

References

Crockett, Robin (2022): Presentation, Properties and Provenance. The Three Ps of Identifying Evidence of Contract Cheating in Student Assignments. In: Sarah Elaine Eaton, Guy J. Curtis, Brenda M. Stoesz, Joseph Clare, Kiata Rundle und Josh Seeland (Hg.): Contract Cheating in Higher Education. Global Perspectives on Theory, Practice, and Policy. Cham: Palgrave Macmillan, S. 169–184.

Microsoft: Rsid Class, Retrieved from:

<https://learn.microsoft.com/en-us/dotnet/api/documentformat.openxml.wordprocessing.rsid?view=openxml-2.8.1> (accessed on 6.6.2023).