Twenty-five years of school? Analysis of Free and Open Source software license texts

Alexios Zavras [a]

(a) Open Source Compliance, Intel Corp.

DOI: 10.5033/ifosslr.v8i1.111

Abstract
The licenses of Free and Open Source Software are expected to be read and understood by all software users. Analysis of these texts shows that it is not an easily achievable goal.[1]

Keywords
Law; information technology; Free and Open Source Software; software licenses; readability

"All men are really most attracted by the beauty of plain speech."
— Henry David Thoreau[2]

Introduction

Free and Open Source software is licensed under a variety of licenses. The text of the license almost always accompanies the software on every delivery and is often included in the software itself. It is generally expected by the software authors and publishers that the users of software will be able to read and understand the software licenses that govern the use of all software and, therefore, subsequently be able to comply with all the license provisions.

The purpose of this paper is to present the results of an automated analysis of the text of the software licenses, which makes it obvious that these expectations that all license texts are read and understood by the users are not easily met.

1 The initial analysis was performed in January 2014; further analysis was completed in the first half of 2016. I am grateful to the participants of the 2014 and 2016 Legal Workshops organised by Free Software Foundation Europe in Barcelona who listened to the presentations and encouraged the publication of this work.

2 Thoreau, H.D., 'A Vigorous Prose Style' in *A Week on the Concord and Merrimack Rivers*, 1849.

Software licenses corpora

While historically, in the early days of Free and Open Source software distribution, many software packages were using their own license text, in recent times this practice has been mostly abandoned in favour of re-using one of a set of commonly used licenses.

The most commonly referenced set of Open Source licenses is that of the licenses that have been approved by the Open Source Initiative (OSI), based on their Open Source Definition. It currently contains 76 licenses.[3]

Since there are many more licenses in wide use than the ones approved by OSI, it was deemed useful to extend the analysis to a wider set of licenses. For this purpose, the entire SPDX License List[4] was selected as a comprehensive corpus of license texts. The results on this paper include the analysis on all 322 licenses present in the version 2.5 of the license list.

It should be noted that the analysis presented here only considered license texts written in English. Although there are licenses in other languages, their use is much more limited. However, a quick analysis of a handful of license texts available in other languages (French and Greek) confirmed that the same general results and corresponding conclusions can be obtained in these cases also.

Presentation of results

The results of the analysis are presented in a series of graphs in the following pages. For each metric, a pair of graphs is presented on a page: the first displays the results for all the OSI-approved licenses, while the second one contains the results for all the licenses in the SPDX license list. To distinguish the OSI-approved licenses in the second graph, they are displayed with a different, darker colour.

In all graphs, the values have been sorted numerically from smallest to largest. This allows viewers to quickly visually recognise the extreme cases, as well as the general pattern of distribution of values. The median value of any metric is the one that is obviously present in the half-way point of the horizontal axis. Each graph also displays the range of values (i.e., the minimum and maximum values) and the average of all the values, rounded to the nearest integer.

Obviously, the position of each license on the horizontal axis does not stay the same, but depends on the metric value for the text of this specific license. It would be an error to assume, for example, that the license that has the maximum value on a specific metric – and thus is placed in the rightmost position – is also on the same position in some other graph displaying the results of another metric.

This paper does not show the exact numerical results for each metric as this depends on a number of assumptions while computing the values. For example, on counting the number of words, one might consider hyphenated words as one or two; or, on counting of sentences, one might ignore – or not – the section headings. However, the general findings are valid independently from such arbitrary decisions.

Moreover, the results are presented in a cumulative and anonymised fashion for all the license texts that were analysed, without detailing or even displaying the exact metric value of each license text.

3 https://opensource.org/licenses/alphabetical
4 https://spdx.org/licenses/

This is because it is not the purpose of this paper to criticize the wording or structure of specific license texts; rather, to raise the issue that all of them share some characteristics.

Basic text metrics

The first set of metrics to be shown are basic quantitative data of the license texts.

The first three pairs of diagrams (fig. 1–6) show the length of license texts, measured in characters, words and sentences. As can easily be seen from these, the length of license texts varies greatly in all metrics. While more than half of the licenses are of a reasonable length, there are some that can be considered extremely long. To give a rough approximation for better understanding of these results, printed books contain between 250 and 300 words per page, while documents such as the papers on this journal have twice that amount.

Having calculated the number of characters, words, and sentences of each license text, the next step is to perform divisions of these numbers in order to calculate averages. The next two pairs of diagrams (fig. 7–10) show this metric: the average number of characters in a word and the average number of words in a sentence. This could give an indication on the complexity of the analysed text. Unfortunately, it turns out that these metrics are not useful and do not provide significant information. The average length of a word does not vary much, and all the values are consistent with reported typical values for arbitrary text written in English. Neither does the average length of sentences, measured in words, vary much – again, it is more a property of the language rather than an attribute of the specific text. There are some outliers with long sentences that appear on the right side of the graph, but these can be explained as licenses with very few – or even a single – sentence, which obviously makes the computation of an average value meaningless.

A metric more interesting than the average lengths is the computation of the maximum lengths, i.e., the longest words and sentences appearing in each license. Obviously, understanding a text presupposes understanding of even its most complex part. The results of the longest words and longest sentences are presented in the next two pairs of diagrams (fig. 11–14). Once again, the results may be somewhat misleading in some cases because of the analysis assumptions. In order to compute the length of the longest word, one has to precisely define what a “word” is. For regular English prose there are only a few decisions to be made – like the aforementioned handling of hyphenated words and counting them as one or two. However, the actual license texts often contain more than regular prose. For example, many of them contain a URL pointing to a resource location; the decision how this should be handled obviously affects the calculated final result. Even if the decision to split the URL into individual path components is taken, it is usually the case that the URL contains lengthy sequences of characters that can be considered words, resulting in a larger number for the length of the longest word.

The metrics presented till now show that even the task of reading the complete license text is, in some cases, not an easy or quick one.

Sentiment analysis

Going beyond the basic metrics of the license texts, the next step is to attempt to analyse the actual content. The last years have seen a remarkable proliferation of algorithms in order to perform

"sentiment analysis" of written text. Sentiment analysis usually computes two metrics for each text: its *polarity* and its *subjectivity*.

The polarity of a text denotes whether the text is positive, negative, or neutral.[5] It is expressed by a number between +1 (most positive) and -1 (most negative), with results close to 0 being most neutral. The polarity of all the license texts are shown in fig. 15 and 16. As can be seen, in both sets, the vast majority of license texts are neutral or slightly positive. There are few exceptions of a few texts being a little negative and a couple of outliers being extremely positive. Thinking about the license contents can provide an insight on the results: licenses usually describe rights and obligations. Such expressions can be formulated in a positive or negative structure; for example, by describing what is allowed to be done or by what is not allowed.

The subjectivity of a text denotes whether the text can be classified as subjective or objective.[6] It is expressed by a number between +1 (very subjective) and 0 (very objective). The subjectivity of all the license texts are shown in fig. 17 and 18. As can be seen, in both sets, the vast majority of license texts are mainly objective. A few exceptions exist also in this case, with some licenses being classified as extremely subjective. This is mainly due to the very small size of these license texts and the presence of words or expressions that might be considered not objective (e.g. "fair").

Linguistic information

The metrics presented above, although giving general information on the license texts, do not provide insight to how easily the license texts can be understood by the people reading them. There is a whole field of research in linguistics that tries to measure exactly this ease of understanding, named *readability* of a text.

There is a plethora of calculations that result in a single number that denotes the readability of a given text. In this paper, results are presented only for a single formula, the SMOG grade. However, the analysis has shown that the results are equivalent when other readability formulas are being used. As with all other metrics, one should be careful with the actual results, mainly because some of the license texts are very short and therefore readability metrics and formulas may not produce an accurate result.

SMOG, which stands for Simple Measure of Gobbledygook, has the advantage of presenting the readability of a text as an estimation of the years of education needed to understand it.[7] For example, a text with a SMOG grade of 6 is deemed to be fully understandable by children having completed 6 years of school.

The SMOG grade of all the license texts are shown in fig. 19 and 20. As can be seen, in both sets, the vast majority of license texts are graded between 15 and 20. A handful of exceptions also exist, with the most sensational being a license that, according to its SMOG grade, can be fully understood by people who have had almost 29 years of school!

5 Pang, Bo; Lee, Lillian; Vaithyanathan, Shivakumar (2002). "Thumbs up? Sentiment Classification using Machine Learning Techniques". *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 79–86.

6 Pang, Bo; Lee, Lillian (2008). "Subjectivity Detection and Opinion Identification". *Opinion Mining and Sentiment Analysis*. Now Publishers Inc.

7 McLaughlin, G. Harry (1969). "SMOG grading: A new readability formula". *Journal of Reading*, 12 (8) pp. 639–646.

Conclusion

The analysis of the language in the licences used for Free and Open Source Software shows that, despite the – stated or presumed – intent of their authors, the licences themselves are not easy reading and cannot be easily understood.

This result is also easily empirically confirmed. There is a vast amount of content available with the sole purpose of explaining the license texts; nevertheless, discussions on the very same subject keep occurring with alarming frequency.

It is reasonable to assume that the writers of these licenses have not purposefully created texts that are difficult to understand; it might even be the case that they have used the simplest possible way to express their intended meaning. However, the undeniable fact is that the current state of the license texts poses a heavy burden on the users.

The actual "cost of understanding" of a license should be always taken into account, especially when endeavouring in the process of creating a new one.

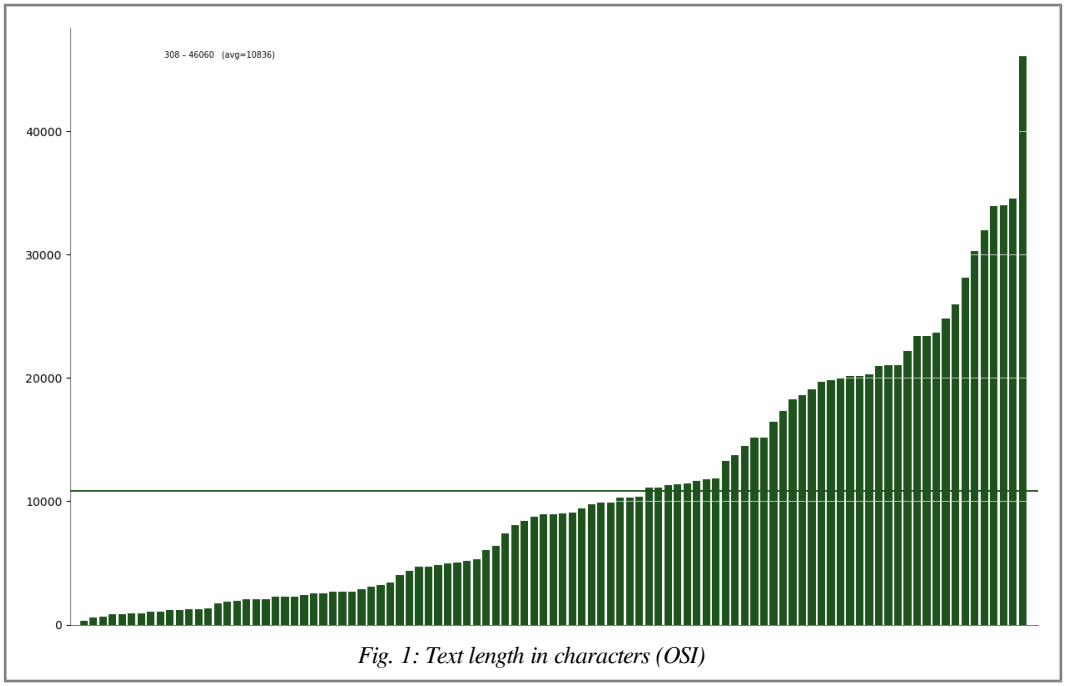

Fig. 1: Text length in characters (OSI)

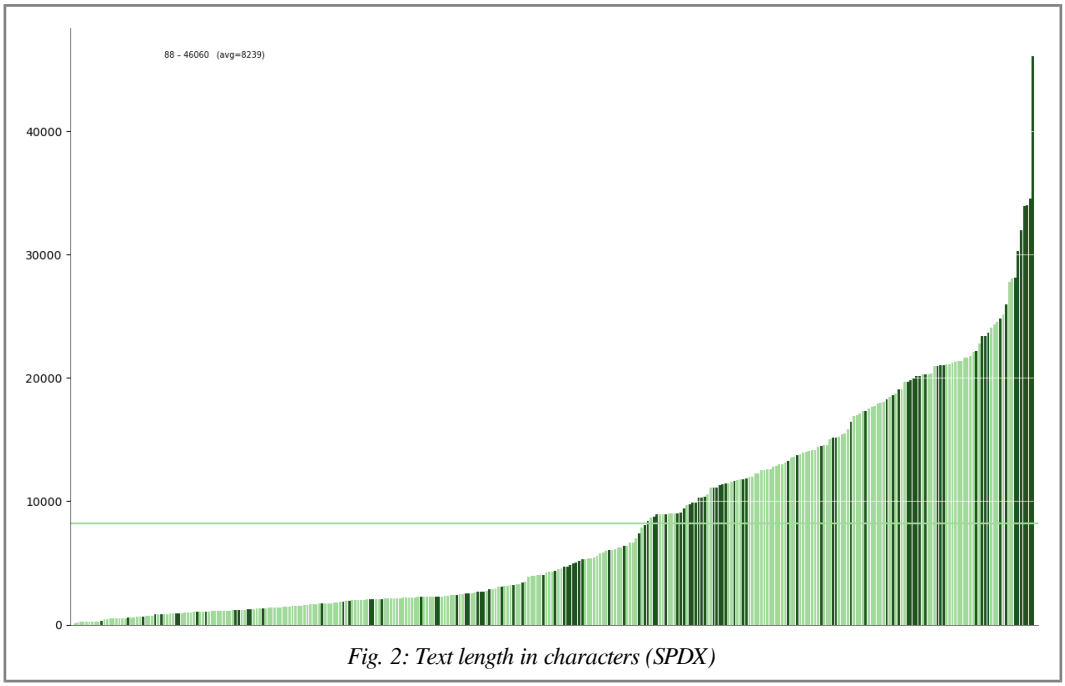

Fig. 2: Text length in characters (SPDX)

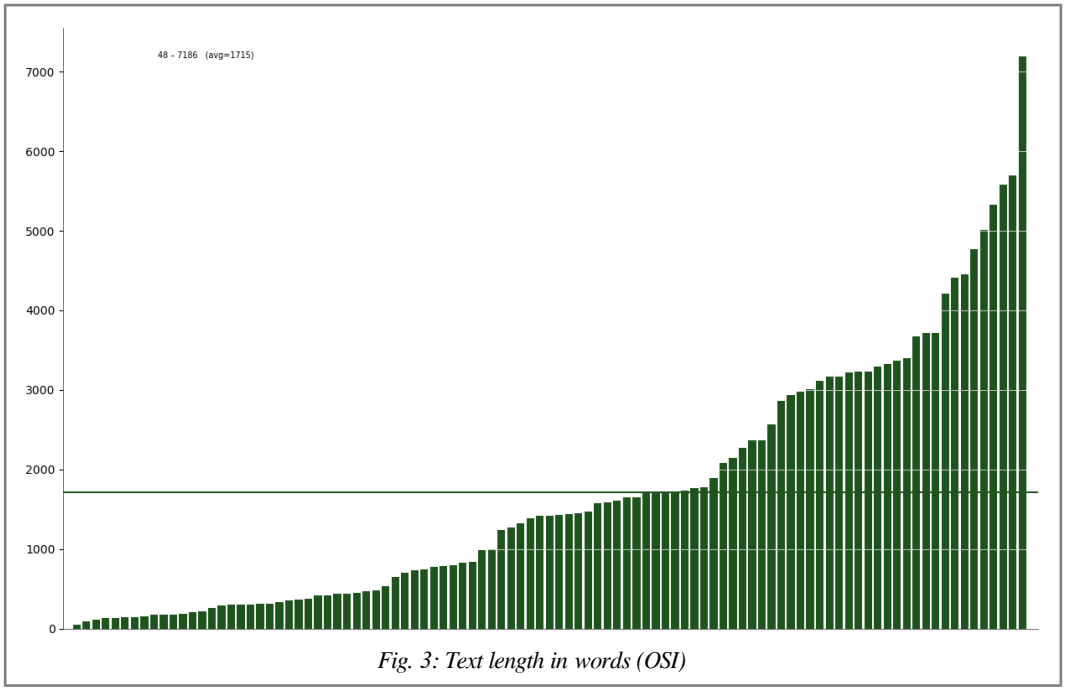

Fig. 3: Text length in words (OSI)

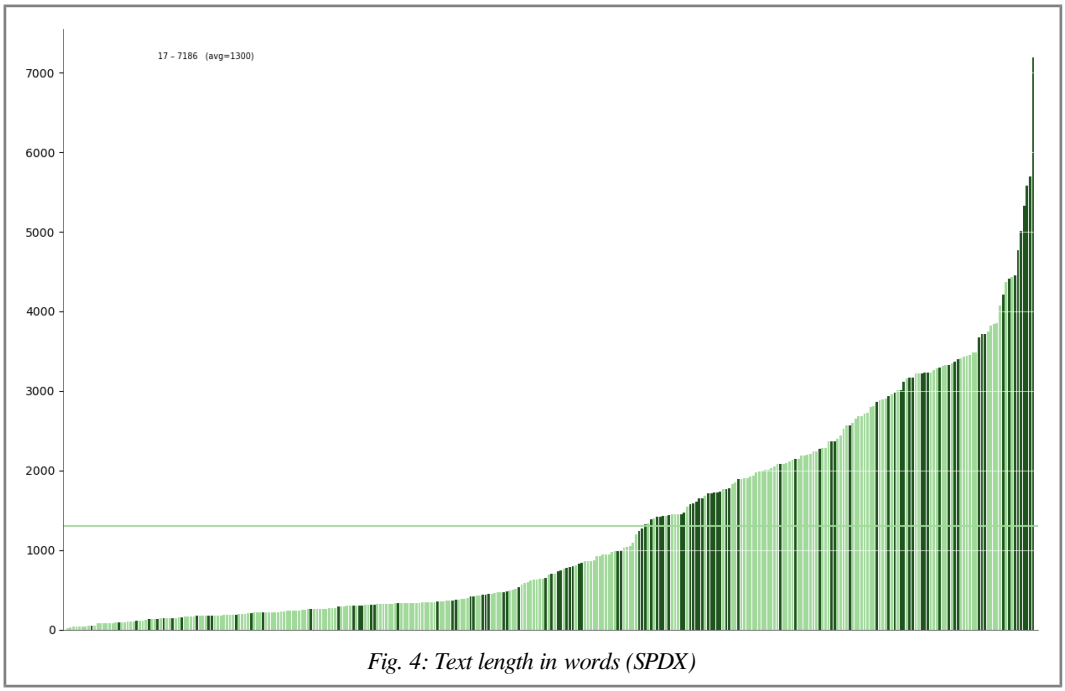

Fig. 4: Text length in words (SPDX)

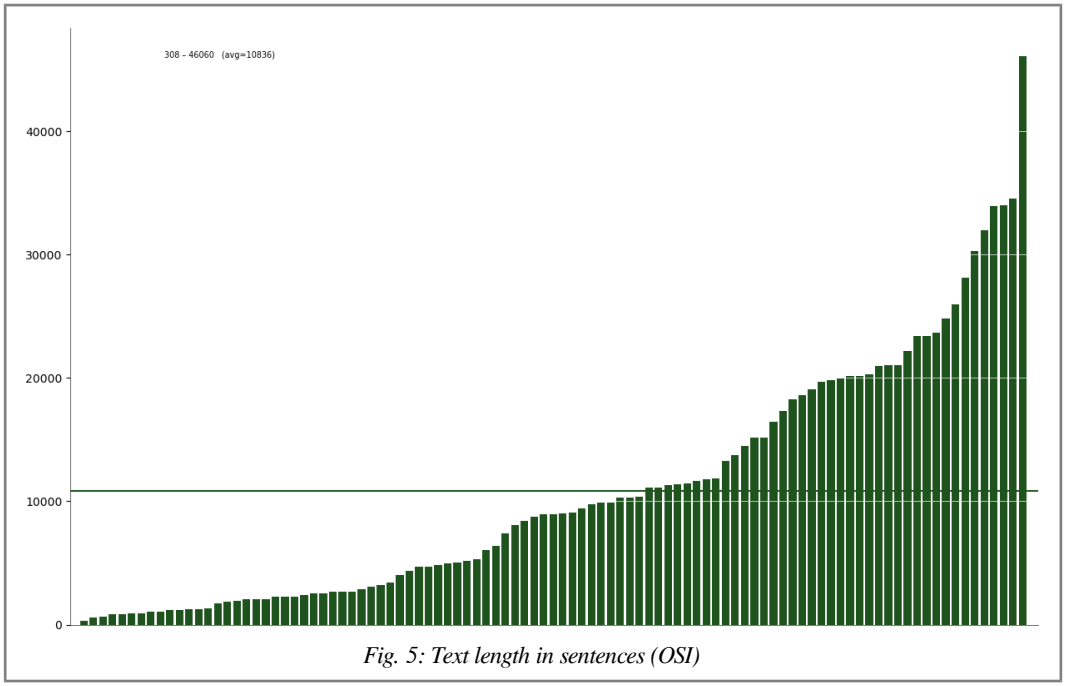

Fig. 5: Text length in sentences (OSI)

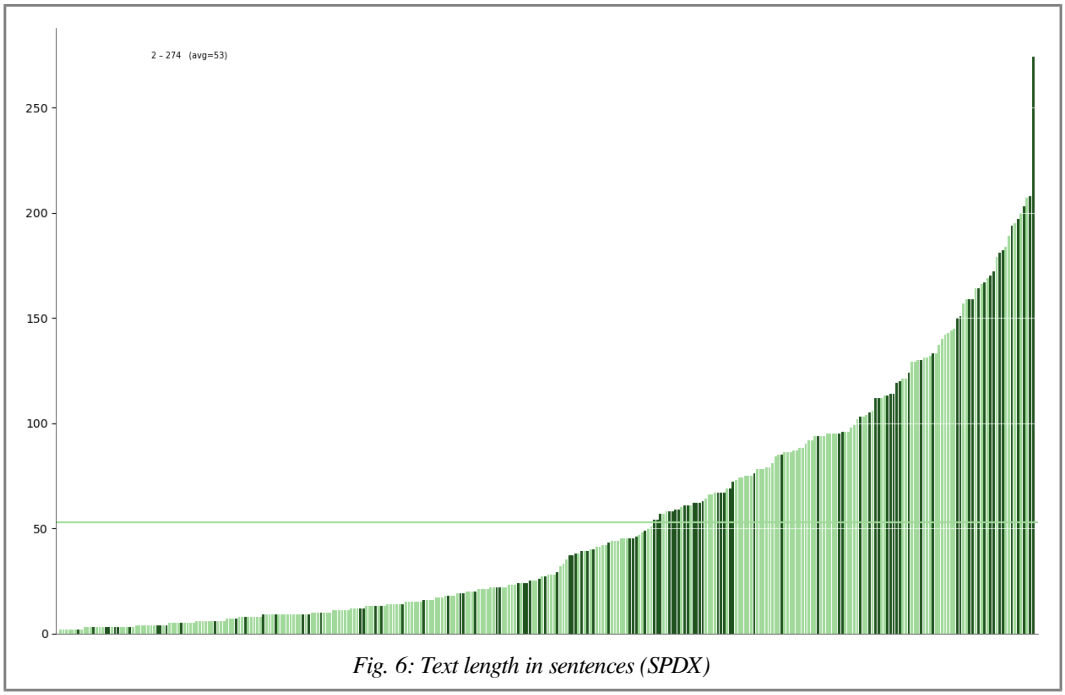

Fig. 6: Text length in sentences (SPDX)

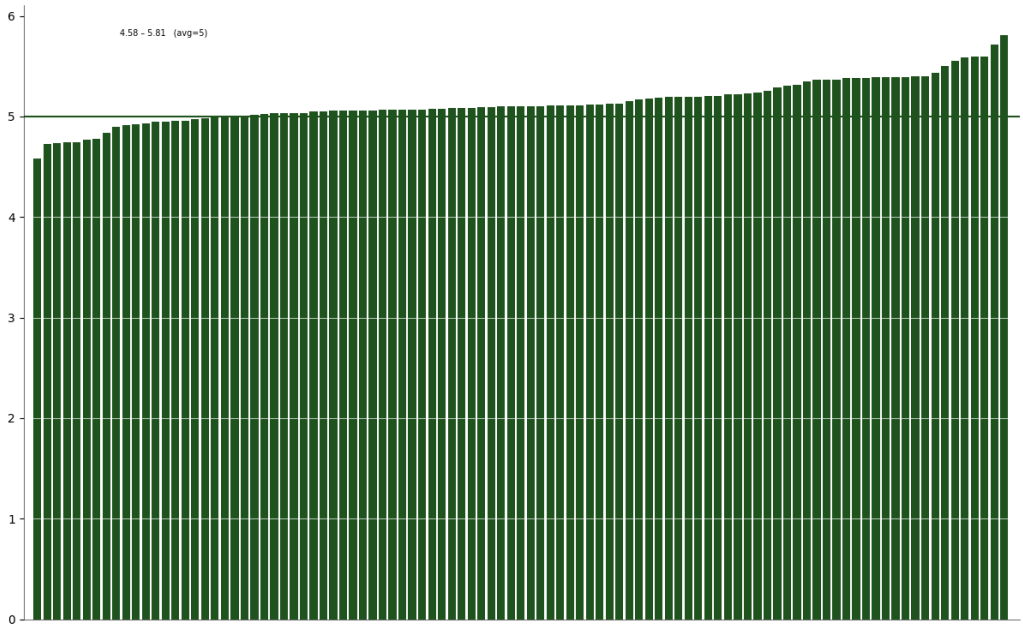

Fig. 7: Average word length in characters (OSI)

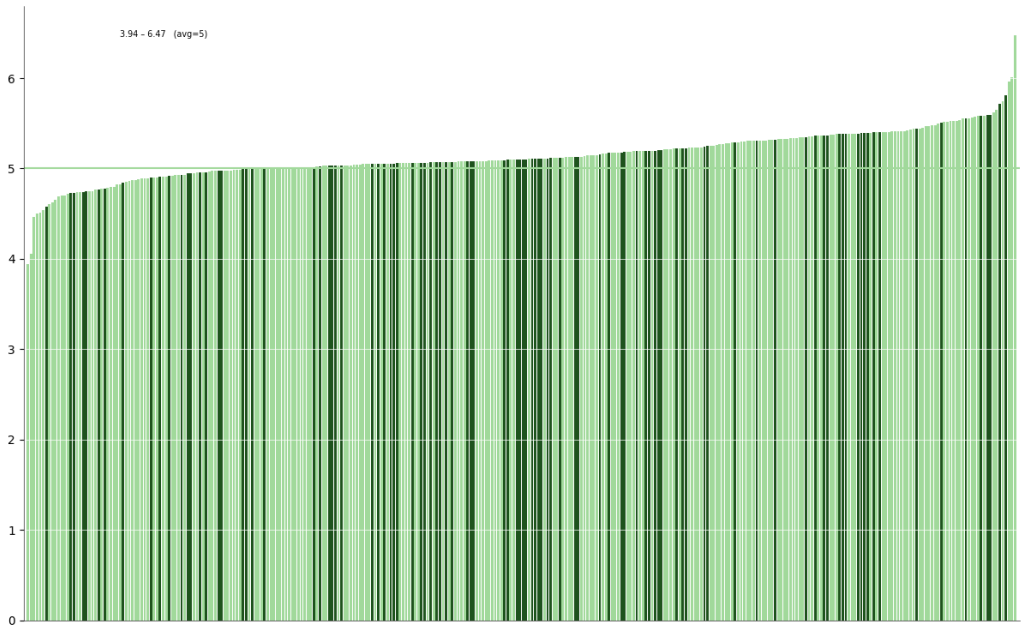

Fig. 8: Average word length in characters (SPDX)

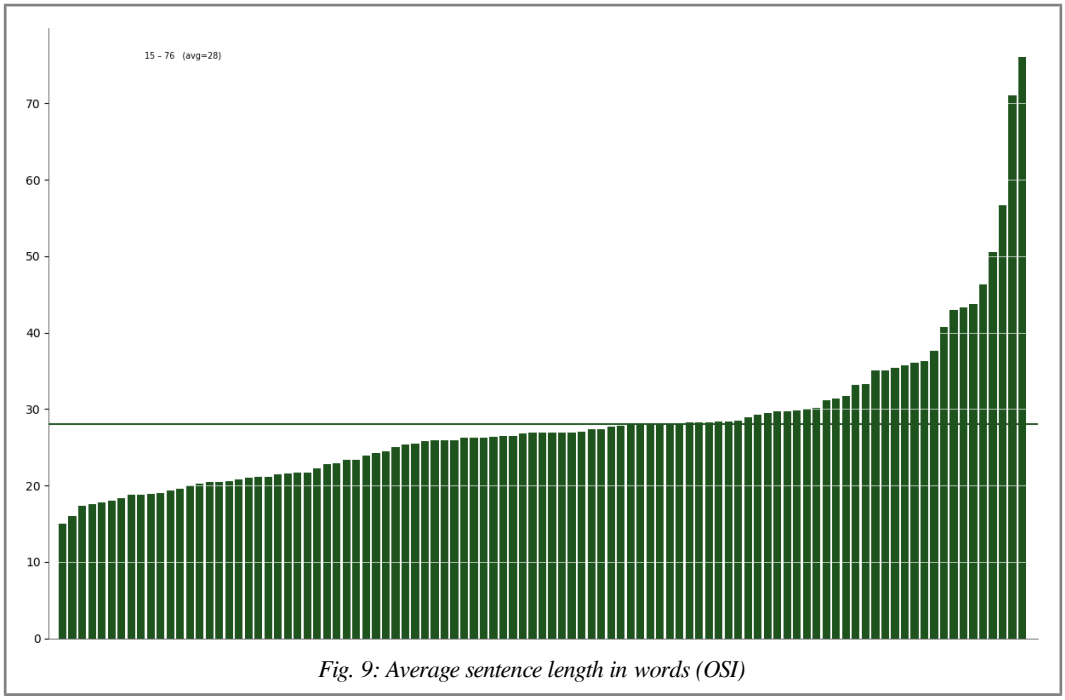

Fig. 9: Average sentence length in words (OSI)

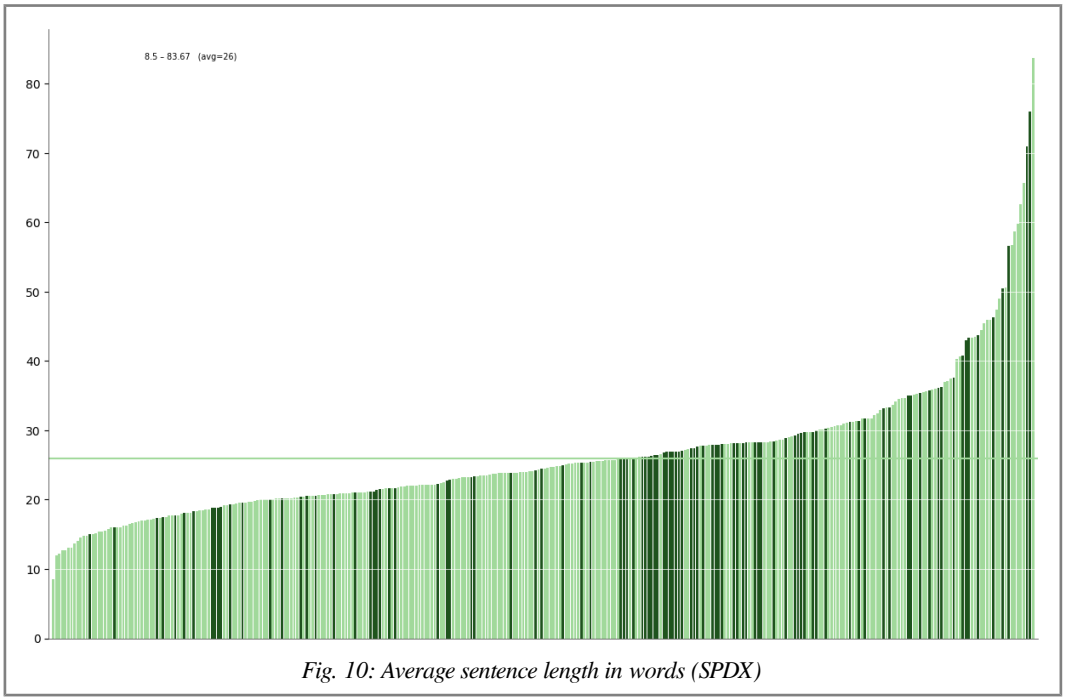

Fig. 10: Average sentence length in words (SPDX)

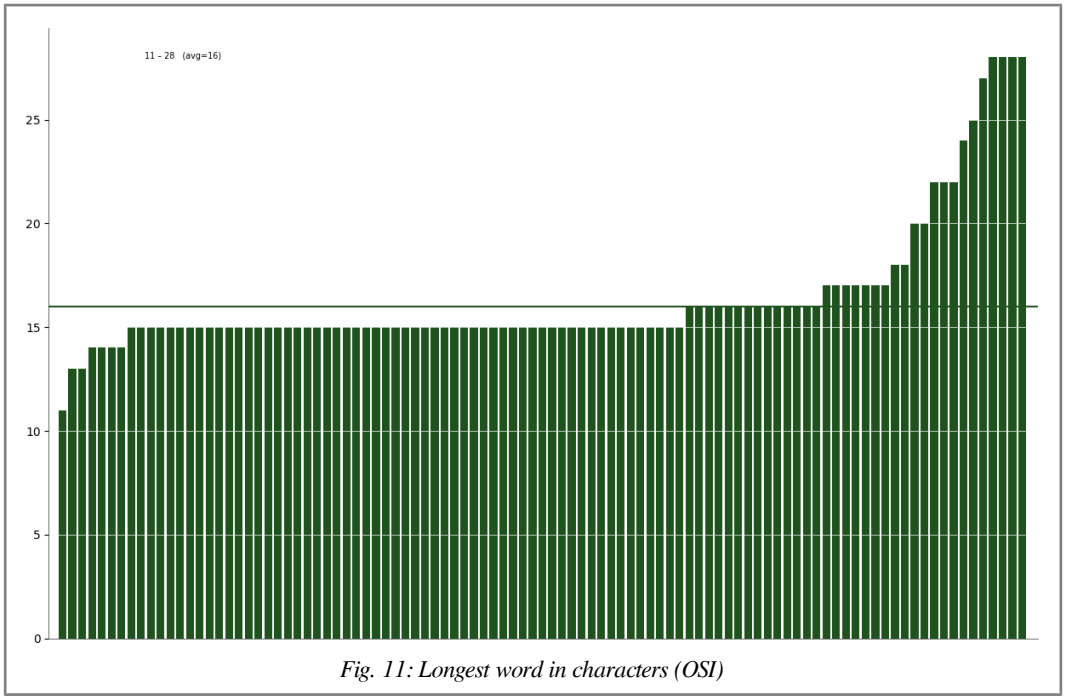

Fig. 11: Longest word in characters (OSI)

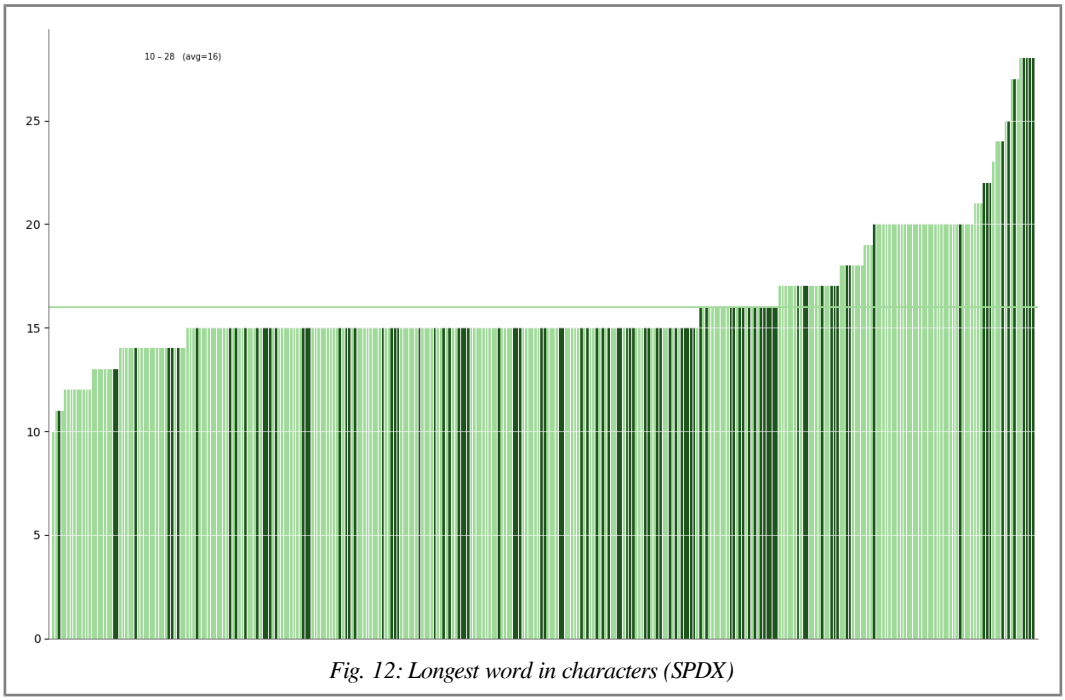

Fig. 12: Longest word in characters (SPDX)

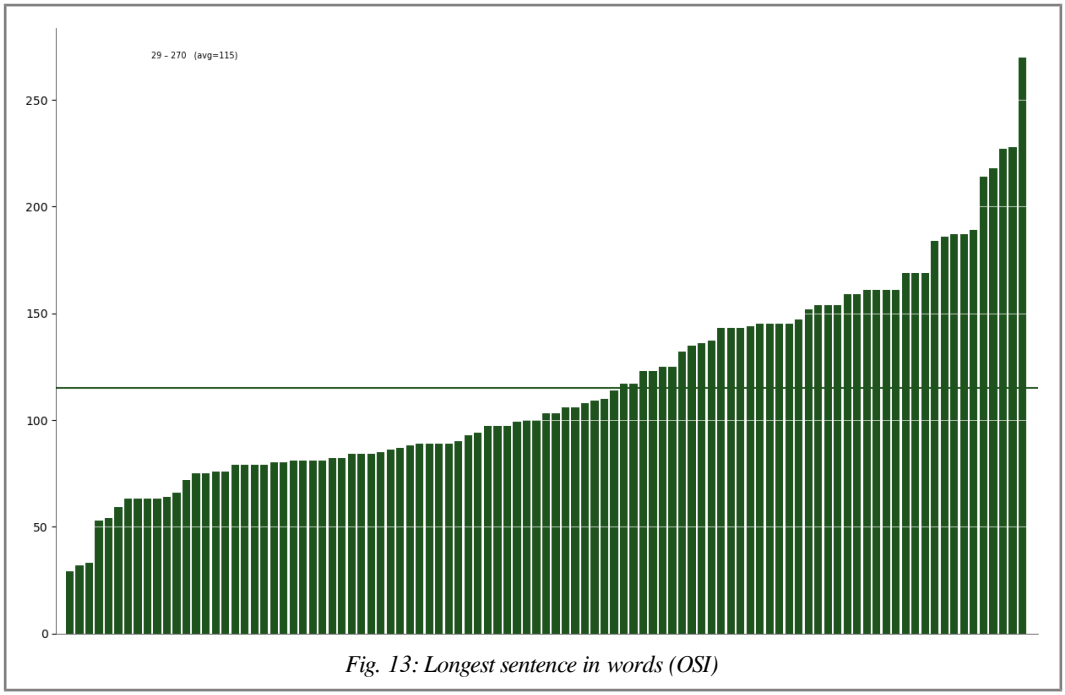

Fig. 13: Longest sentence in words (OSI)

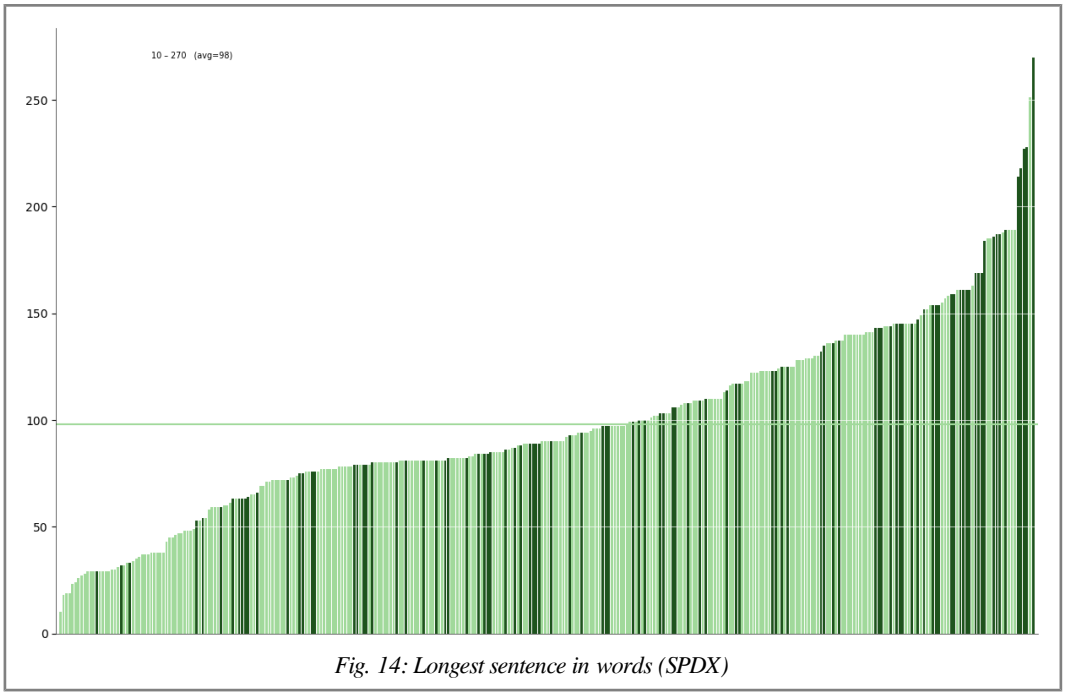

Fig. 14: Longest sentence in words (SPDX)

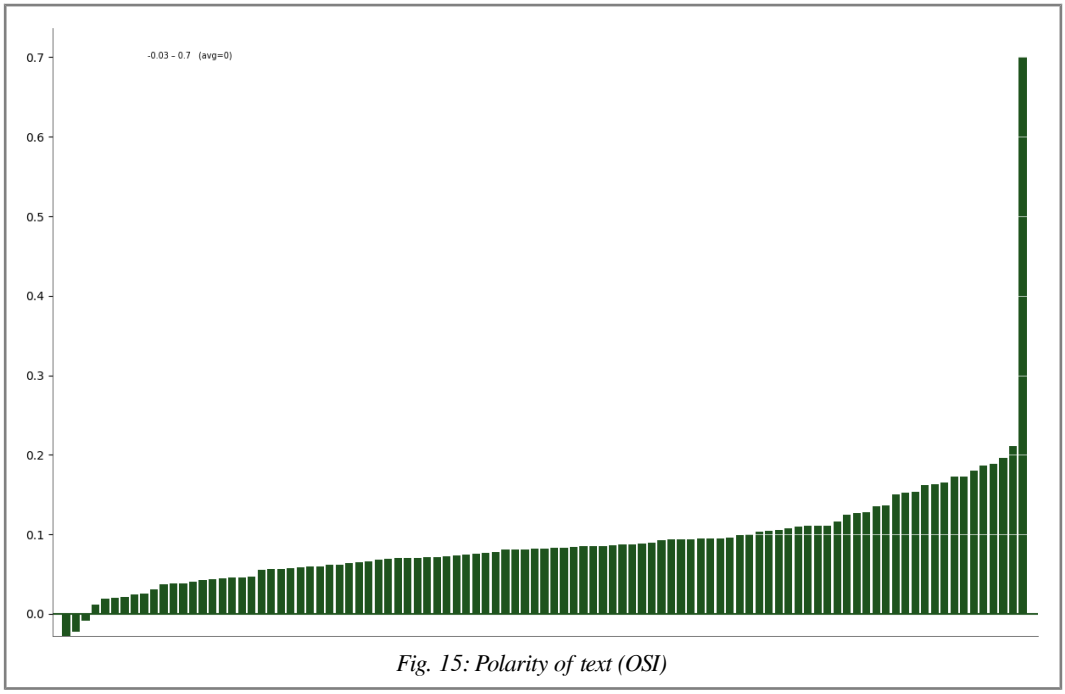

Fig. 15: Polarity of text (OSI)

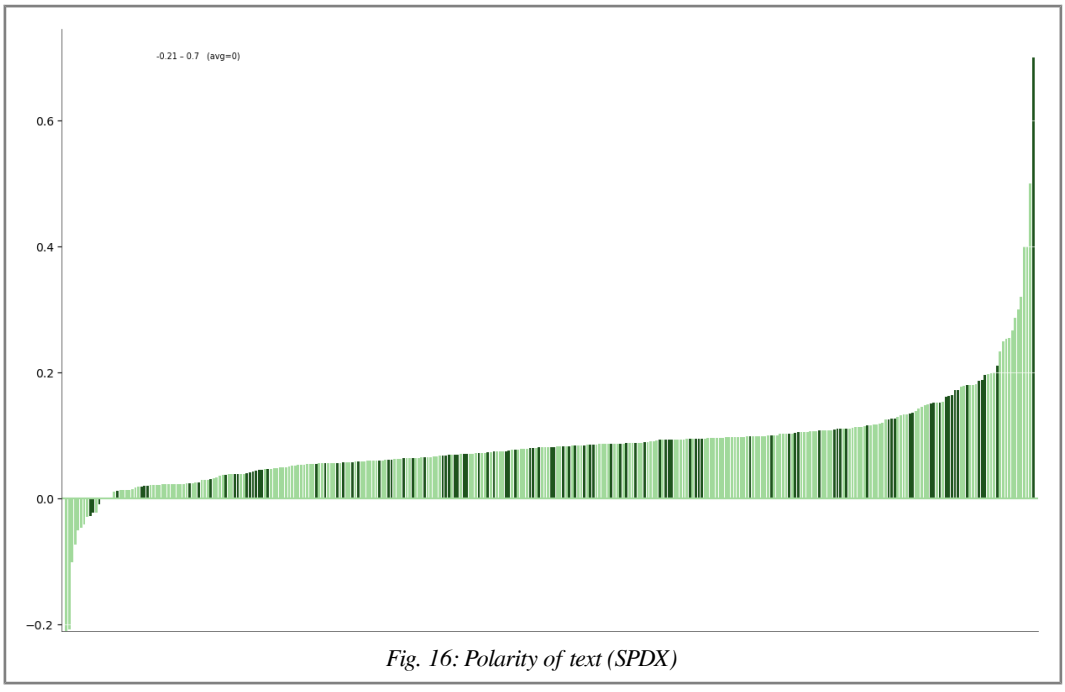

Fig. 16: Polarity of text (SPDX)

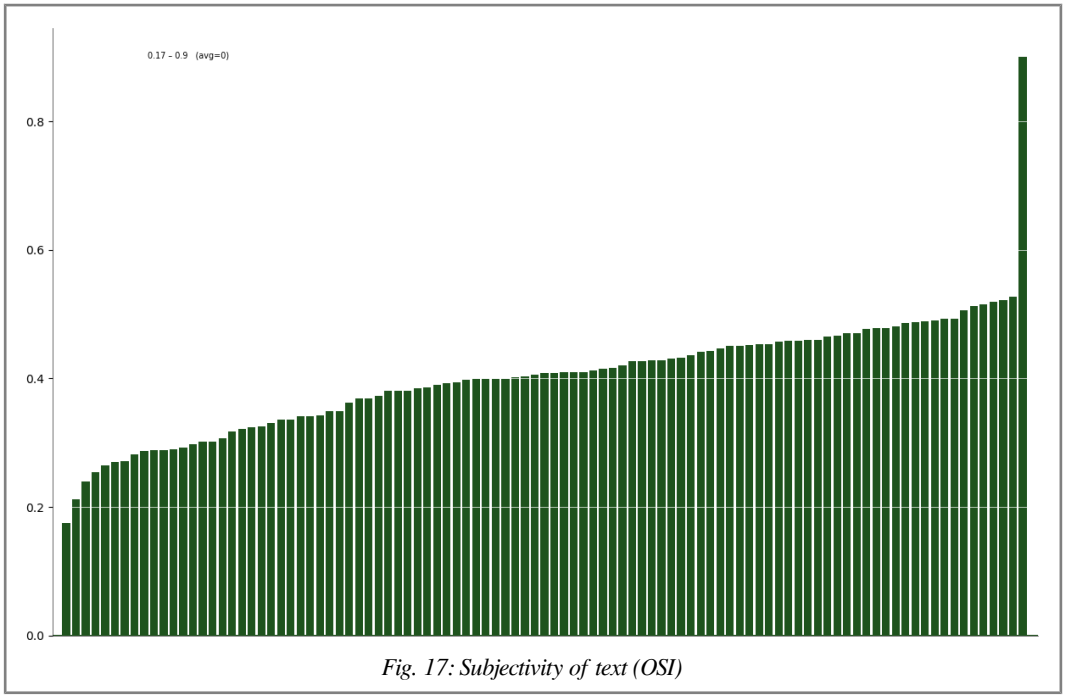

Fig. 17: Subjectivity of text (OSI)

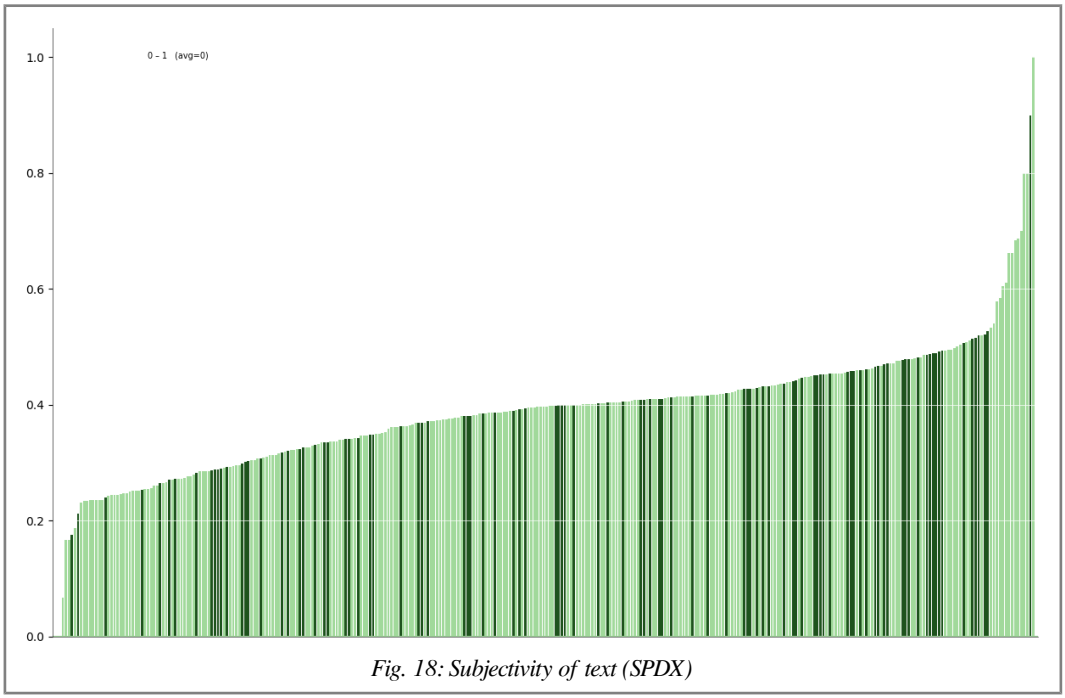

Fig. 18: Subjectivity of text (SPDX)

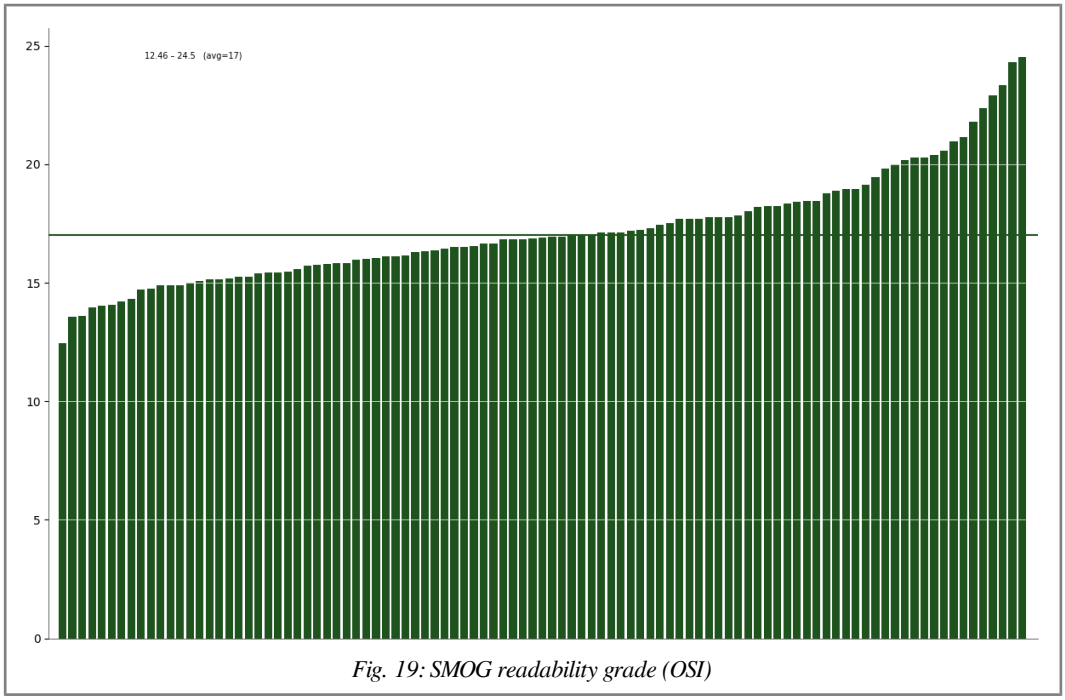

Fig. 19: SMOG readability grade (OSI)

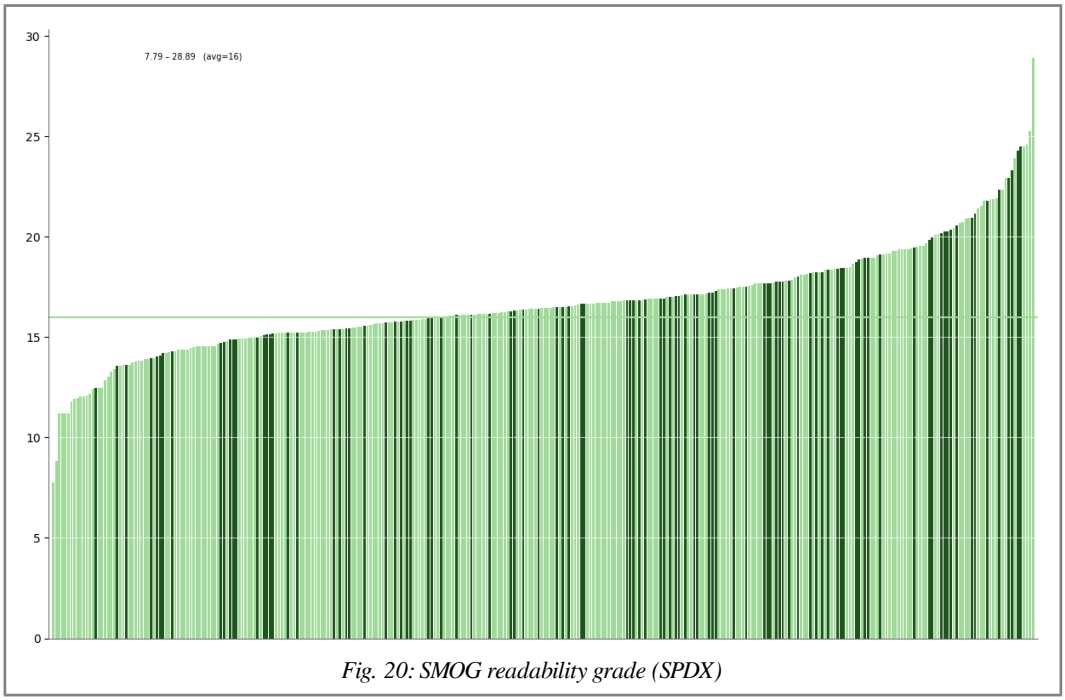

Fig. 20: SMOG readability grade (SPDX)

About the author

Alexios Zavras *is the Senior Open Source Compliance Engineer of Intel Corp. He has been involved with Free and Open Source Software since 1983, and is an evangelist for all things Open. He has a PhD in Computer Science after having studied Electrical Engineering and Computer Science in Greece and the United States.*