FAIRness of documents released under the Dutch Freedom of Information Act

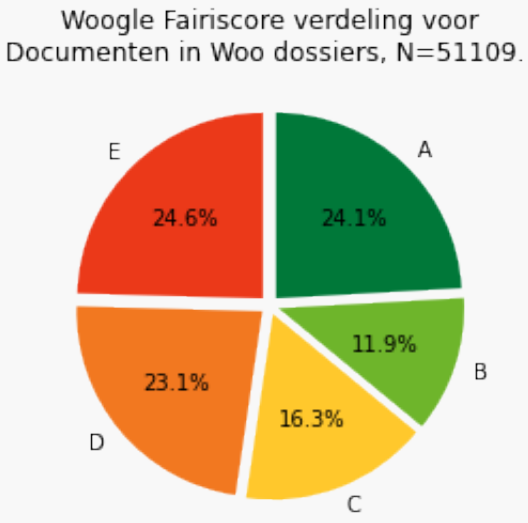
- Maarten Marx, Jaap Kamps, Informatics Institute and ILLC, University of Amsterdam
- maartenmarx@uva.nl, kamps@uva.nl

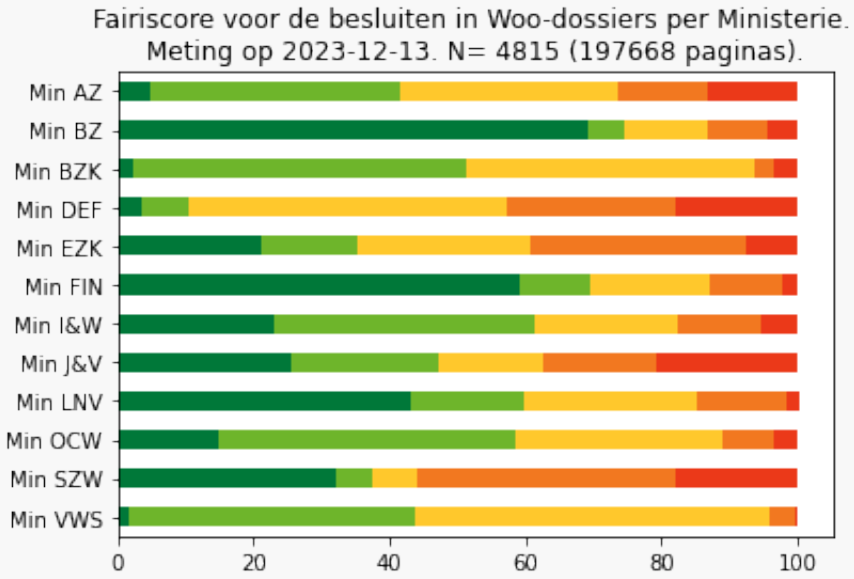
The Dutch Freedom of Information Act, the *Wet open overheid (Woo)* is in full force since May 1st 2022. Documents released under this act have to be (according to the act) machine readable, contain relevant metadata and be compliant with European accessibility and re-use guidelines. We have measured to what extent this is the case, using the FAIR scientific data principles[1] as yardstick. The results are in general not very good.

Our research on this topic has three dimensions:

1. measuring the problem and creating awareness;
2. repairing existing documents, turning them into FAIR data;
3. preventing new bad quality documents by improved software and workflows.

For measuring, creating awareness and providing an incentive to change, we developed the *FAIRiscore*, a five point scale indicating how similar the human and the machine readable text of a document are. The piechart below describes the distribution of the scores over 51K documents containing more than 1.2 million pages. Documents with score E do not contain any machine readable text at all, only documents with score A are most likely digital born, all others are scans with varying text quality mostly due to OCR errors. The stacked barplot shows how this score varies for documents published by the various ministries. This indicates that different workflows are being used and the good examples, like *Min BZ* (Foreign Affairs), show that large improvements are possible.





For repairing, we measure the compliance to the Web Accessibility Guidelines WCAG and are able to repair the three most pressing problems with accurate results: no or bad OCR, absence of crucial metadata, absence of semantic tags.

Finally, for preventing these unneeded problems, we have shown that text redaction (which happens a lot in FoIA documents) can be done efficiently and safely in digital born PDFs. There is no need to use the, in The Netherlands very common, scan and OCR method.

Woogle

We apply these results in the search engine for Dutch FoIA documents, Woogle. The developed software is made available via https://wooverheid.nl/ and will be presented at the poster session.

1. Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18. ↩