Case description from the project “FAIRify Humanities” of the National Data management Forum 2019.

2019-12-03

Title: ensiwiki-2011 dataset for readability modelling

Name and affiliation: Frans van der Sluis / HUM-KOMM-IBID, KU.

Research area: Detecting readability and textual complexity from text. This is all data from my PhD work.

Data set(s): The first dataset contains 40,000 "mature" Wikipedia articles, of which 20,000 from the "simple english" and 20,000 from "english" Wikipedias. The second dataset contains 18 articles from The Guardian and subjective ratings about their complexity as provided through a user study. This combination of datasets provides a valuable benchmark for training and testing textual readability and complexity models. It provides for a comparison between objective performance (on The Wikipedias) and subjective ratings (on The Guardian) as well as for an evaluation on two distinct genres of content (encyclopedic vs. news articles). The first plan was to make the first dataset available via CLARIN-DK. However, it was necessary to postpone the final submission of the dataset as it was chosen that publishing the Wiki-data combined with the Guardian-data together would optimise the potential impact.

What efforts was carried out to make your data more FAIR?

More “Findable”: The data will be described with metadata in CLARIN-DK. A PID will be available for the data. And a readme file will be added describing the data.

More “Accessible”: Through CLARIN-DK, the data and metadata will be made accessible according to contemporary standards.

More “Interoperable”: The readme file is formatted in plain text using simple Markdown formatting. The data itself were offered as SqLite3 file, which assures the correct data types are used and makes the relations between the different data tables evident operable.

More “Re-usable”: A readme file is added with instructions on how to access the data file. The usage license was kept the same as Wikipedia’s, CC-BY-SA 4.0. The original data was formatted as Wiki text, but also a plain text version will be added to reduce the required pre-processing for potential users and increase the comparability of research based on the dataset.

What was the biggest challenges to make your data more FAIR?

1. Copyright

Both the Wikipedia data and The Guardian data are copyrighted. Wikipedia releases its data under a creative commons license which allows for reuse and redistribution. The Guardian releases its articles openly, but does not allow for redistribution.

Finding out about the different licenses turned out more difficult than expected. For the case of Wikipedia, conditions were stated on the redistribution of their data. For the case of The Guardian, finding the right contact point and engaging in negotiations proved difficult. In either case, copyright documents are not trivial to read and interpret. This indicates a need for expertise on copyright and funds for releasing copyright.

2. Added value and sensitivity

It was hard to estimate what the added value of releasing data is. Yet, this was necessary to decide on which data to release. Similarly, the sensitivity of data was also hard to estimate whilst it determined what can(not) be published. In practice, these challenges resulted in a precautious stance on what to publish. For example, to publish summary data (ie. aggregated) rather than individual data points that could be traced back to individual participants who joined the experiment. Some guidelines and best practices might lead to more informed and less precautious decisions about what can and should be published.

3. Reliability and comparability

An issue in this field of research is the use of openly available yet copyrighted documents for training and testing algorithms. Given that the used corpora are copyrighted means they cannot be redistributed, which in practice makes that each researcher collects their own selection of documents from public sources also used by other researchers. In addition, different researchers employ different pre-processing to the data they obtain. As a result, even though studies often use the same source of documents, their results are not necessarily comparable. This points to a challenge with the reliability of data. To allow for proper reuse and comparable results requires to resolve copyright issues and release pre-processed versions of the data.