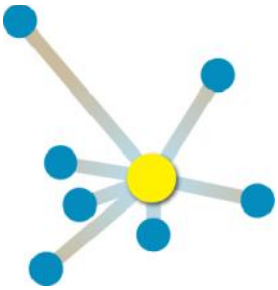
European Language Resource Coordination (ELRC) is a service contract operating under the EU's Connecting Europe Facility SMART 2015/1091 programme.



European Language Resource Coordination

Connecting Europe Facility

Comment on the 6th LRB Meeting:

Review of K. Heafield's presentation from a Legal Perspective

Author(s): Pawel Kamocki (ELDA)

Dissemination Level: Public

Version No.: <V1.1>

Date: 2018-05-16

Following the presentation of ParaCrawl, Kenneth Heafield presented several questions on legal issues (in particular personal data and copyright issues) which are reviewed below from a legal perspective:

- First of all, he pointed out that the European Language Resource Association (ELRA) sells webcrawled parallel corpora which, of course is a normal procedure provided that ELRA has obtained the rights for distributing the LR. He then referred to a text snippet of a resource that said "Or check the video done by Steve Huff" and said that some of these resources are not anonymized. ➔ *NB: The lack of anonymization in this case naturally doesn't present a problem because "Steve Huff" here is the name of the author of a video, and mentioning the name of the author whenever a work is used in fact is a legal obligation.*) Also, Kenneth Heafield showed the extract of a copyright notice in this resource stating that "All information published on this website is copyright protected and may not be used without written permission from Schaper & Brümmer." (*NB: This is correct and the resource is still available because the provider warrants that he obtained all necessary permissions. As such, it is correct to provide the resource in this way.*)
- Following this, Kenneth Heafield provided the example of the National Library of the Netherlands, saying that The Netherlands has no legal deposit law and that the National Library archives the web anyway. He further suggested to adopt a pragmatic way to handle copyright issues: The opt-out approach which assumes implicit permission for web archiving. ➔ *NB: It is very important to note that there is a fundamental difference between web archiving by a national library (which is based on the Dutch transposition of the library exception, art. 5.2(c) and 5.3(n) of the InfoSoc Directive) and a body like ELDA who distribute LR commercially, the actions performed within a funded project like ELRC or ParaCrawl (which includes the provision of services). ParaCrawl resources (like the resources collected by ELDA and ELRC) were not compiled by a publicly accessible library, so no parallel can be drawn.*
- Kenneth Heafield also claimed that "[e]very EU case which was cited in [ELDA's legal evaluation] report was won by a crawler". ➔ *NB: This unfortunately does not correspond to the reality. Quite the contrary: e.g. Infopaq and Directmedia were quite obviously 'against' crawling. It shall also be remembered that the Court of Justice of the European Union only interprets law and does not apply it to the facts of specific cases (this is later done by national courts), so it cannot be said that a CJEU case was 'lost' or 'won' by anyone.*
- Kenneth Heafield moved on further to illustrate different quotes from ELDA's legal evaluation report, namely: "not impossible to organize the crawling process in such a way as to comply" vs. "the creation of statistical language models would also qualify as lawful use". ➔ *NB: Here it is important to point out that these quotes were taken out of their context. Because yes, it is not impossible to structure the*

crawling process in such a way to comply with copyright law, providing that no permanent copies of the crawled content are made, and that there is no communication to the public. ParaCrawl on the other hand obviously made permanent copies and is now communicating them to the public (see tinyurl.com/ybtda3dk) which may naturally cause problems with regard to copyright law. And also, yes, it is very likely that the creation of language models will qualify as lawful use – it is one of the five elements of the exception for temporary acts of reproduction that need to be met cumulatively. ParaCrawl clearly does not meet the first one (see above), so arguing about the remaining four is an exercise in futility.

- Kenneth Heafield also displayed to further quotes from ELDA's legal evaluation report, to show that it apparently ignores options and jumps from "it seems that the most viable way" to "only the sources that pass this validation procedure". ➔ *NB: Again, the quotes are out of context because the report actually says: "**The most viable way** of making sure that the crawling operations are lawful is to perform an a priori clearance of the sources that are to be crawled. (…) [In this approach], **only the sources that pass this validation procedure can be lawfully crawled".***
- In the following, Kenneth Heafield suggests as option for the use of crawled resources the ELRC-endorsed exception for temporary acts of reproduction. ➔ *NB: The information about the possibility to apply the exception for temporary acts of reproduction is summarized as follows: "Unfortunately, these exceptions allow for web crawling only in very limited circumstances. This is the case when: the reproductions made in the process are temporary (which is of very limited relevance for crawling activities) (…)" Here it is very important to note that the temporary character of the reproductions is a sine qua non for the application of the abovementioned exception (called rather descriptively: the exception for <u>temporary</u> acts of reproduction). Once again, reproductions made by ParaCrawl are not temporary (they are not a part of the process in which they are automatically deleted) so the exception cannot apply, regardless of whether other conditions are met.*
- As a saner approach, Kenneth Heafield suggests to use ROAM (Randomise – shuffle the sentences, Omit – remove data, Anonymise – replace all personal data, Mix – jumble sentences from different sources). ➔ *NB: From a legal point, the ROAM approach is by no means saner for a number of reasons:*
    - *<u>Randomise</u>: shuffle the sentences -- in order to shuffle the sentences one needs to download them all first, i.e. make reproductions; as explained in the report, this in most cases requires permission from the rightholder (which may be granted up front, e.g. in a public license, hence the importance of a priori IPR clearance); moreover, to anticipate the counterargument, the exception for temporary acts of reproduction does not allow for modifications of the reproduced works;*

- o *Omit: yes, in Germany up to 15% of a work can be communicated to the public for research and teaching purposes only, only in a password-protected environment, and subject to the payment of compensation to a German collecting society – does ParaCrawl meet the remaining requirements? Is ParaCrawl really omitting 85% of the data?*
- o *Anonymise: removing phone numbers and e-mail addresses is far from enough for robust anonymisation; still, anonymization using appropriate techniques is the way to go.*
- o *Mix: "jumbling sentences" means making derivative works which requires permission from the right holder. Moreover, in many jurisdictions (e.g. in France) it is likely to be regarded as violation of a moral right of integrity (protecting works against distortion*).
- ⇨ *As such, legally, the ROAM process doesn't make it legal to obtain the LR – it just makes it more difficult to get caught in the end.*

- Furthermore, Kenneth Heafield moves towards case law referenced in the ELDA's legal evaluation report. He claims that Google image search provides thumbnails of images from the web and that uploading on the open internet (without any measures to prevent indexing) would yield an implied consent (Vorschaubilder I), even if uploaded without the rightholders permission (Vorschaubilder II). He explains that linking can infringe copyright only if the user "knew or ought to have known" that the links leads to infringing content, and that commercial providers are presumed to have this knowledge. In Vorschaubilder III, however, no presumption of knowledge was applied against Google, despite its being commercial. This, in contrast to Kenneth Heafield, does not mean that the doctrine of implied consent has been abandoned, as the ELDA report argues. ➔ *NB: The presentation slides again misquoted ELDA's legal evaluation report; the report does not say that the German court "**abandoned** implied consent", it says: "[the court] **seems to have abandoned** (…) implied consent". There is a fundamental difference between "abandoned" and "seems to have abandoned". Also, it is important to note that the German case Vorschaubilder III is indeed based on CJEU's decisions in Svensson and GS Media cases (which seemingly means that previous German cases: Vorschaubilder I and II have been overruled), but it distinguishes between the facts in these two CJEU cases. In short, the German court ruled that because Google is a search engine provider, the presumption established by the GS Media case (that providers of commercial websites are aware of the illegal nature of the content that they link to) does not apply. Therefore, the solution of Svensson (linking is not communication to the public) prevails. The facts of the Vorschaubilder cases are of very limited importance for web crawling (unless performed by search engines providers), this is why the report only mentions them very briefly. Vorschaubilder III is essentially a case about linking, not about crawling data. As a consequence, even if the implied consent doctrine survived in Germany, it seems that it only applies to openly available search engines – and not to*

crawling. Until we know of any recent case law that applies this doctrine to other scenarios, we cannot reasonably and responsibly advice anyone to base its crawling activities on implied license/consent. In most countries implied license/consent is not conceivable, as the law requires that licenses be in writing.