

The future of BootCaT:
A Creative Commons License Filter

egon w. stemle & verena lyding

<{egon.stemle, verena.lyding}@eurac.edu>

Institute for Specialised Communication and Multilingualism



European Academy of Bozen/Bolzano
(EURAC)

EURAC
research

June 24th, 2013

"Copyright issues remain a gray area in compiling and distributing Web corpora"

William H. Fletcher, 2011

Notwithstanding all the good intentions of a researcher who collects web data for building a corpus in the name of *fair use* (i.e. without committing an act of piracy), redistributing data taken from the web without the permission of their creator is – strictly speaking – illegal.

Marco Brunello, 2009 (adapted)

... solved?

"If a Web corpus is infringing copyright, then it is merely doing on a small scale what search engines such as Google are doing on a colossal scale"

Adam Kilgarriff, Gregory Grefenstette, 2003

"Starting August 1, Google News in Germany will only index sources that have decided to explicitly opt-in to being shown on the search giant's news-aggregation service. Google News remains an opt-out service in the other 60 countries and languages it currently operates in, but since Germany passed a new copyright law earlier this year that takes effect on August 1, the company is in danger of having to pay newspapers, blogs and other publishers for the right to show even *snippets* of news"

TechCrunch, June 21st, 2013

The problem (with the "Leistungsschutzrecht" Law) is that it is not clear when *a few words* become *a snippet*.

... solved?

"If you want your webpage to be removed from our corpora, please contact us"

<http://wacky.sslmit.unibo.it/doku.php?id=corpora>

"Even if the concrete legal threats are probably minor, they may have negative impact on fund-raising"

Anke Lüdeling, Stefan Evert, Marco Baroni, 2007

Adding the possibility for minimizing the legal risks, or rather, actively facing and eliminating them is paramount to the WaCky initiative.

Documentation

GNU Free Documentation License, Apple's Common Documentation License, Version 1.0, FreeBSD Documentation License, Open Publication License, Version 1.0, Open Content License, Version 1.0

Works of Practical Use Besides Software and Documentation

GNU General Public License, GNU Free Documentation License, (some) Creative Commons Licenses, Version 2.0, Design Science License (DSL), Free Art License, Open Database license

Works Stating a Viewpoint (e.g., Opinion or Testimony)

GNU Verbatim Copying License, Creative Commons Attribution-NoDerivs License, Version 3.0

Standardised Way to Grant Permissions by Mixing Conditions

(by) Attribution: Licensees may copy, distribute, display and perform the work and make derivative works based on it only if they give the author or licensor the credits in the manner specified by these.



(sa) Share-alike: Licensees may distribute derivative works only under a license identical to the license that governs the original work.



(nc) Noncommercial: Licensees may copy, distribute, display, and perform the work and make derivative works based on it only for noncommercial purposes.



(nd) No Derivative Works.



http://en.wikipedia.org/wiki/Creative_Commons_license

16 Possibilities between "all rights reserved" and "no rights reserved" – 11 Valid and 6 Regularly Used Licenses:

- Attribution alone (by)
- Attribution + NoDerivatives (by-nd)
- Attribution + ShareAlike (by-sa)
- Attribution + Noncommercial (by-nc)
- Attribution + Noncommercial + NoDerivatives (by-nc-nd)
- Attribution + Noncommercial + ShareAlike (by-nc-sa)

Marking Content with CC Licensing Information

Insert HTML code into the webpage so that your work is clearly marked.

3 steps to license notice perfection:

- The full URI (link) to the license. Example:
`http://creativecommons.org/licenses/by/3.0/us/`.
- A visible notation (most commonly text) that states the license being used.
- Optionally, the appropriate Creative Commons license button or CC icon and license property icon(s).

This work is licensed under a

```
<a rel="license"  
  href="http://creativecommons.org/licenses/by/3.0/deed.en_US">  
  Creative Commons Attribution 3.0 Unported License  
</a>.
```

Platform for Corpus-Assisted Italian Language Learning

One part of the corpus was constructed using a method inspired by the WaCky project:

50,000 randomly combined words from an Italian basic vocabulary list were used to retrieve candidate pages with the Yahoo! search engine.

Hits were limited to certain Creative Commons licenses: by, by-sa, by-nc-sa, and by-nc. Pages that were wrongly tagged as CC-licensed were eliminated using a black-list that was populated by manual inspection of earlier versions of the corpus.

The retrieved pages were automatically cleaned using the KrdWrd system.

Simple&Stupid `grep 'creativecommons.org/licenses'` yields:

From 200,534 CC-licensed web pages from the PAISÀ corpus all but 1060 were identified as containing a CC license link (99.95%).

From 10,000 randomly selected non-CC-licensed pages from a crawl of Italian web pages 15 were wrongly identified as containing a CC license link (0.15%).

i.e. identification of CC-licensed pages similar to Yahoo!

	Trentino	Trentino (cc_any)
URLs	4495	
Sites	1655	
(ratio)	2.7 : 1	
URLs	4500	4426
Sites	2422	449
(ratio)	1.9 : 1	9.9 : 1
Shared Sites	35	
blog URLs	267	797
blog Sites	211	98
(ratio)	1.3 : 1	8.1 : 1
Shared blog Sites	3	

Table: Two rounds of BootCaT-ing to construct a specialised corpus about the Trentino region. Common first round; second round with identical seed terms *and* search queries - without and with CC restriction. But cf. Brunello, 2009: The creation of free linguistic corpora from the web.