EUROPROwebs Labelling and Tagging

The verbal and visual components of the H2020 websites selected were downloaded and saved into plain text documents. The verbal texts were specifically stored as txt-documents, so that they could be processed and worked upon with quantitative software tools (e.g. AntConc, Wordsmith, #LancsBox; UAM Corpus Tool, SketchEngine).

A master document was saved to include the whole content of the website (labelled NAMEOFTHEPROJECT). Individual documents were also saved corresponding to the most common and salient web sections or pages (e.g. Homepage; About us; Partners; Contact). Unifying labels were agreed to group the sections and pages from the different project websites, regardless of their original name in the website menu in question.

Table 1 below illustrates the labels that we chose to name the documents of the EUROPROwebs Digital Corpus, as well as the diversity of original names that web sections and pages were given by the international research projects:

Labels to name the sections of international project websites	
HOME	Home; Homepage; Ø
ABOUT	About; About the project; About Us; Mission; Project; Concept
PARTNERS	Partners; Consortium; Team; Researchers
NEWS	News; Events; News and Events; Blog; Newsletter
WORK	Work Packages; Demo Sites; How we work; Model
OUT	Output; Publications; Materials; Media; Outlets; Guidelines; Deliverables
NET	Networking; Participate; Get involved; Sister projects

In order to organize and store the documents from the Internet, specific labels were chosen to indicate the prominence and place of the corresponding sections within the international project website they are housed in. Table 2 highlights the three main options we came across in our corpus:

Place of the section within the website menu	
MAIN	The website genre appears explicitly in the web menu as a visible top label that users can click on and interact with.
SUB (subsumed)	The website genre is inserted within the web menu, but the user needs to hover the mouse to unfold a further menu and click on the desired option.
EMB (embedded)	The website genre or section is occluded within one of the menu options, so the user needs to navigate the web to arrive there.

As regards the tagging of the EUROPROwebs corpus, this was carried out at two different levels.

1. **Manual tagging**: A number of tags was determined by the research group in the light of the texts explored and download. Three main foci were of interest to us:

   a) Interactive hyperlinks: <int> internal; <per> peripheral; <ext> external

   b) Multimodal elements:
   <images>, <video>, <carrousel>
   <maps>, <table>, <graph>, <questionnaire>, <presentation>

   c) Languages of use: English is the fundamental language of the corpus and is used as an International Language or as a Lingua Franca. Tags were included for Spanish (<sp>) and the rest of languages altogether (<lang>).

   The whole corpus was thus annotated.

Projects

<link>LEARN MORE ABOUT OUR PROJECTS</link>

The MIGRATE Members

The Transmission System Operator (TSO) TenneT is the project coordinator of MIGRATE. Together with 23 other partners (10 TSOs, 12 Universities/Labs and 1 manufacturer) of the consortium the research and investigation will be done.

<link>LEARN MORE ABOUT THE MIGRATE MEMBERS</link>

News and Newsletter

Here you will find recent news about MIGRATE and the possibility to register for our newsletter.

<LINK>VIEW MORE</LINK>

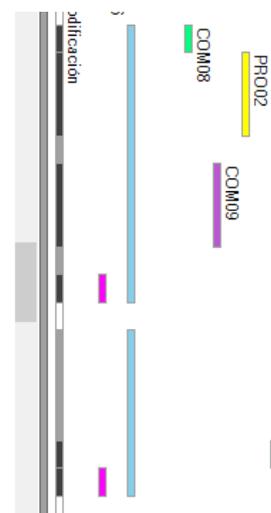Example of manual tagging in EUROPROwebs Digital Corpus

2. **Semi-automatic tagging**: Depending on the research aims targeted and methodological perspectives followed, different tagging at the lexico-grammatical, discursive or pragmatic level was devised using the software for qualitative analyses NVivo12.

Cefic (European Chemical Industry Council) is the Brussels-based organisation representing the European chemical industry. Cefic represents, directly or indirectly, about 29,000 large, medium and small chemical companies, which employ about 1.2 million people and account for nearly a third of world chemical production. Cefic was incorporated in 1972 as an international association with scientific objectives. Cefic addresses key issues of global importance, such as: Competitiveness, Research & Innovation, High Level Group on Chemicals, EU Chemicals Policy, International Trade, Climate change & Energy, etc.
- Martin Winter: mwi@cefic.be
- <link>http://www.cefic.org/</link><ext>

CEMBUREAU is the European Cement Association based in Brussels and represents national cement industry associations and cement companies from the European Union (except Cyprus, Malta and Slovenia) plus Norway and Switzerland, before European and national governments.
- Koen Coppenholle: k.coppenholle@cembureau.eu
- <link>https://cembureau.eu/</link><ext>

Example of semi-automatic tagging in EUROPROwebs Digital Corpus

Related publications:

Pascual, D., Mur-Dueñas, P. and Lorés, R. 2020. Looking into international research groups' digital discursive practices: Criteria and methodological steps in the compilation of the *EUROPRO* digital corpus. *Research in Corpus Linguistics 8* (2), 87-102. https://doi.org/10.32714/ricl.08.02.05

To visit the EUROPRO digital Database click HERE