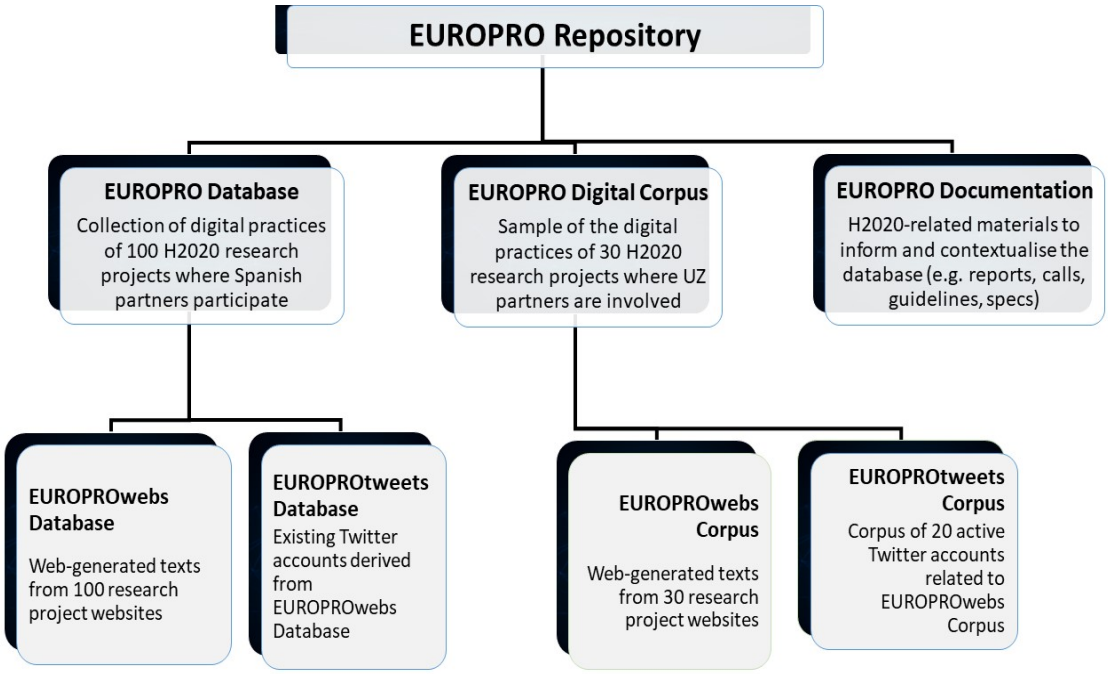
EUROPRO Digital Corpus Description

The **EUROPRO Digital Corpus** is a collection of digital texts that reflect the discursive and linguistic practices of international research projects financed by the Horizon2020 European framework for excellence science and research innovation. As such, it is regarded as a specialized corpus of texts that is static, implying that it was compiled at one specific point in time. The EUROPRO Digital Corpus is divided into two sub-corpora focusing on specific objects of inquiry, namely the **EUROPROwebs Corpus** (European Project Websites) and the **EUROPROtweets Corpus** (European Project Twitter Accounts). Overall, the corpus allows us to analyse digital academic discourse and, through its analysis, insights can be gained into international research group's evolving discursive and professional practices.

Additionally, as the figure below displays, the EUROPRO Digital Corpus is part of a holistic network of genres, texts and documents that we have labelled as **EUROPRO Repository** and that go deeper into the academic, professional and contextual aspects that surround international research projects. Thus, this repository comprises the **EUROPRO Database**, which is an extension of the EUROPRO Digital Corpus and include a total of 100 Horizon2020 international research projects. Much as the EUROPRO Digital Corpus, two collections of texts make up this database: the EUROPROwebs Database containing 100 research project websites and the EUROPROtweets Database collecting the existing Twitter accounts derived from projects in the database. Lastly, the EUROPRO Repository has also a branch devoted to **EUROPRO Documentation**, where materials and resources are gathered to gain understanding into how international research projects are created and developed and how scientific knowledge is expected to be communicated, disseminated and exploited.



Going back to the EUROPRO Digital Corpus, which is our main object of study, now we described the two sub-corpora mentioned above, and explain the methodological choices we made in order to compile it.

1) EUROPROwebs Corpus (European Project Websites) contains the web-generated and web-hosted texts and documents of **30 research project websites** created and maintained by international research groups. Different web sections are therefore part of this corpus: *Home, About, Partners, News and Events, Blog, Work Packages, Publications, Deliverables* or *Contact*, among others. EUROPROwebs Corpus contains 394 072 words with a mean of 13 136 words per website.

2) EUROPROtweets Corpus (European Project Twitter Accounts) was compiled as an extension of EUROPROwebs Corpus, considering the importance of social media, and Twitter in particular[1]. It includes the 20 Twitter accounts that were found to be active out of the 30 research projects whose websites make up the EUROPROwebs Corpus. Tweets, retweets, replies and threads were collected as part of the feed of tweets published by international research projects.

EUROPROtweets pilot consists of 3 822 tweets containing 88 970 words, with a mean of 191 tweets and 4 449 words per research project Twitter account.

Criteria for Corpus Compilation

Several criteria were followed for the selection of the 30 international Horizon2020 funded research projects that are comprised in the EUROPRO Digital Corpus.

1. The research projects should aim at producing new knowledge and not at training PhD students or professionals to undertake the research. This was done to ensure comparability within the corpus.

2. A convenience sampling method was followed, which entailed choosing research projects with at least a member from the Universidad de Zaragoza (Spain) or from a Zaragoza-based institution. This would allow us to complement our text-based analysis with contextual evidence from potential informants.

[1] The use of Twitter is encouraged by institutional guidelines, together with other social networks, and was found to be the preferred option for dissemination purposes by H2020-funded research projects in our corpus. Check our Deliverable EUROPROtweets Corpus Description to find out more details.

3. The date of the H2020 projects should at some point overlap with the development of our own project (2017-2021), so that the most recent digital academic practices could be studied.

4. Given their organic nature, the websites were selected and downloaded on a fixed date, which would entail that the duration of each of the projects was different at the moment of compilation. Since creating a monitor corpus (a corpus that is regularly updated) lies beyond the scope of our research project, websites had been developed to different stages when downloaded.

Methodological Decisions taken in EUROPRO Digital Corpus

Four key notions derived from the digital nature of the EUROPRO Digital Corpus were taken into account prior to its compilation, i.e. 'dynamicity' and 'hypermodality/hypermediality', as common to the digital texts of both EUROPROwebs and the EUROPROtweets corpora, and then 'layout and web design' in the case of EUROPROwebs corpus and 'interactivity' in the case of EUROPROtweets corpus. Specific details about our considerations for these four problemtic aspects are offered as follows.

EUROPROwebs corpus

All texts from the websites were downloaded and labelled using different codes to refer to the pages or sections of the website.

1. **DYNAMICITY** The corpus was compiled from April to May 2019. The specific start and end date of each project was retrieved in order to determine the point of development of the corresponding research project at the moment of the corpus compilation.

2. **HYPERMODALITY AND HYPERMEDIALITY**. The corpus was tagged for hyperlinks (external, internal and peripheral), for visuals (such as tables, figures, pictures, logos, etc.), for videos and audios.

3. **LAYOUT AND WEB DESIGN**. Besides downloading the texts contained in the website, screenshots of every page were taken and saved. Information was also recorded of the extent to which the text could be directly accessed from the website menu.

EUROPROtweets corpus

1. **DYNAMICITY**. All tweets and retweets from research project Twitter accounts were coded and downloaded at a set date, June 2019.

2. **HYPERMODALITY AND HYPERMEDIALITY**. Information regarding multimodal elements and hyperlinks in the tweets was coded. Regarding multimodality, three multimodal elements were found and, accordingly, coded: pictures, videos and GIFs.

3. **INTERACTIVITY**. Information was retrieved on the number of likes obtained by each tweet, and the number of retweets by other users. Information was also retrieved and coded on the number and range of hashtags (#) used in the Tweets by the research group and the number and range of mentions (@) to other Twitter users.

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 Corpus click HERE