Enriching Europeana Newspapers

Enriching Europeana Newspapers aims to expose the full text aggregated as part of the Europeana Newspapers project. It contains over 11 million pages of full text of historic newspapers (mainly but not all 19th century), drawn from national and research libraries across Europe. A portal is already in place. This pilot aims to expose and improve the text for more data driven usage (ie large scale data analysis of the whole corpus).

The Scientific Challenge

Creating best practice guidelines for the publication, citation and impact measurement of cultural heritage data (ie the newspapers in question). Standards for citing and judging the impact of open cultural data are still far from being established.

Enriching the newspapers corpus, via the automatic extraction of topics and named entities; the current corpus is only searchable via free text searches.

Showcasing the value of the enrichment by a quantitative analysis of the occurrence of topics/entities over time and across borders. A particular challenge will be the extraction of topics across texts in multiple languages (over 40 languages are featured in the corpus from French to Yiddish to Estonian) and variable quality of the digitised text. Digital humanities scholars will be interested in the raw OCR texts; the number of these is likely to be in the 100s rather than 1000s. We also suspect that others in linguistics, economics, information science and computer science can make use of the datasets If successful the enriched texts could also be placed in the current Newspapers interface (http://www.theeuropeanlibrary.org/tel4/newspapers). This received over 1.4m page impressions in 2015, around 5 to 6,000 users a month. Better search facilities will help improve these numbers.

Who benefits and how?

The publication of the data will benefit a wide variety of research communities. Ongoing work has already identified interest in the data from the following communities:

- A broad spectrum of historians, or other digital humanities disciplines where topics of interest are included in the newspapers (economic historians, art historians, literary historians, geographers…).
- Linguists, given the linguistic diversity of the source (over 40 languages featured) it will be immensely valuable.
- Computer and information scientists working in the areas of natural language processing, information retrieval, information extraction, text mining and optical character recognition.

Standards for the citation of cultural heritage data in research are much required. Citing such data is not just about providing a stable URL, but also providing the kind of documentation that allows scholars to assess the quality and scope of the data. The latter is especially important when dealing with corpora of mixed-quality data: original data (that may have been produced by different OCR engines and having differing levels of fidelity to the original), manual annotations and automatically created enrichments. Cultural heritage institutions that are making such data available also need such standards. Additionally, they require better metrics and processes to

EUDAT receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654065.

understand the benefits of making such data openly available.

In their current form, the newspapers can be accessed through their metadata fields or by full-text search. It is currently not possible, however, to group newspaper articles based on topics. An enrichment of article-level annotations with named entities and topics will benefit researchers wishing to select a portion of the corpus that is relevant to the topic of their study. In addition, the enrichment will enable a quantitative study of the topics and entities discussed in the news.

We aim to increase awareness among a wide range of scholars of the content of the created dataset as well as methods to access and process it. For that purpose, we will perform a quantitative analysis of a selection of the enriched collection to visualize which were the 'hot topics' in the selected period and how this was different across different countries. This will demonstrate to scholars working with the corpus what kind of data and metadata it contains, and how this kind of quantitative analysis can be done, paving the way to more quantitative studies on the newspaper corpus in the future. The aim is to open up discussions about the value of working with big data as well as the problems of dealing with errors in the data (e.g. those caused by OCR or enrichment tools).

In their current form, the newspapers can only be accessed via a web portal, thus computer scientists are unable to obtain the full corpus for machine processing. The availability of the corpus through the EUDAT infrastructure will enable further data and text processing to be applied and derived research datasets to be created. These will further promote the reuse of the corpus by several research communities.

Technical Implementation

Europeana is using the B2SAFE and B2FIND services to help undertake the enrichment of the datasets and, more generally, expose them for re-use by other academics, particularly those outside the digital humanities.

Contacts

- Nuno Freire, Europeana Foundation/The European Library, nfreire(at)gmail.com
- Maciej Brzezniak, PSNC, maciekb(at)man.poznan.pl

Further Information

Scientific Challenges behind the Pilot
To learn more about the EUDAT/Europeana collaboration visit here

Category:
Data Pilot Communities

Read more

EUDAT receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654065.