Enabling language processing of cultural heritage content in Digital Humanities: interoperability between Europeana and CLARIN

Maria Eskevich (CLARIN ERIC), Twan Goosen (CLARIN ERIC), Nuno Freire (Europeana / INESC-ID), Clemens Neudecker (Berlin State Library / Europeana Newspapers)

To realise the full potential of available large-scale cultural heritage datasets and to stimulate the use of such data through data-driven research practices in the field of digital humanities, we need to enable interoperability of digital media content with processing tools.

Collaboration between Europeana [1], a European cultural heritage (CH) infrastructure, with CLARIN [2], Europe's Common Language Resources and Technology Infrastructure, is addressing the implementation of such a goal, and provides use cases that can be shared with the community to stimulate wider uptake.

Overall, by nurturing partnerships that place the data directly in the workflow of the researcher, and that allow the retrieval of high-quality reusable CH data, it becomes possible to reduce the transaction costs for finding and mining data resources for research purposes.

In this paper we describe the technical aspects of research infrastructures integration, and focus on the evaluation of the interoperability between cultural heritage data and the language tools accessible via the CLARIN metadata-based portal for language resources: the Virtual Language Observatory [3].

Europeana seeks to enable users to search and access knowledge in all the languages of Europe, either directly via its web portals, or indirectly via third-party applications leveraging its data services. One of the lines of Europeana's actions, is to facilitate research on the digitised content of Europe's galleries, libraries, archives and museums, with a particular emphasis on digital humanities. The Europeana service is based on the aggregation and exploitation of (meta)data about digital objects, using the Europeana Data Model (EDM) as its model for metadata interoperability [4]. EDM was initially defined with discoverability use cases in mind, but in its latest updates, it also includes the metadata to indicate the technical details of media content, which enable its processing by language tools.

CLARIN is a networked federation of language data repositories, service centres and centres of expertise. CLARIN aggregates metadata from resource providers (CLARIN centres and selected "external" parties), and makes the underlying resources discoverable through the Virtual Language Observatory (VLO) to provide a uniform experience and consistent workflow.

The VLO can also serve as a springboard to carry out natural language processing tasks via the Language Resource Switchboard (LRS), allowing researchers to invoke tools with the selected resources directly from its user interface. The inclusion of many new CH resources by 'harvesting' metadata from Europeana, has increased the potential for new application scenarios based on CLARIN's processing tools.

Components from both infrastructures had to be extended or adapted to accommodate the integration process [5, 6]. Currently, about 775 thousand Europeana records can be found in the VLO, from which 10,000 are technically suitable for processing via the LRS. This amount is less than 1% of the complete Europeana dataset, therefore, our ongoing work aims to greatly increase these amounts.

In parallel, we carry out an investigation into the overall coverage of the integrated material in terms of processability, and provide estimates for the amount of resources and tools in different languages that can be both discovered in the VLO and processed with LRS. This evaluation provides insights for researchers in the humanities and social sciences into potential ways to use the Europeana's high-quality data, and apply a wide range of digital research methods.

References

[1] http://www.europeana.eu

[2] www.clarin.eu

[3] https://vlo.clarin.eu

[4] Europeana (2016). Definition of the Europeana Data Model.
https://pro.europeana.eu/resources/standardization-tools/edm-documentation

[5] T. Goosen, CLARIN and Europeana: Cultural Heritage Data for the Digital Humanities. CLARIN Annual Conference, Budapest, Hungary, 2017.
https://www.clarin.eu/sites/default/files/clarin2017-bazaar-cultural-heritage-data-for-dh.pdf

[6] T. Goosen, N. Freire, C. Neudecker, M. Eskevich. Bringing Europeana and CLARIN together: Dissemination and exploitation of cultural heritage data in a research infrastructure. Digital Infrastructures for research (DI4R), Brussels, Belgium, 2017.
https://indico.egi.eu/indico/event/3455/session/1/contribution/14