Science Demonstrators' Success Stories

Author(s)	Franco Niccolucci
Date	25/2/2018
Science Demonstrator Title	TEXTCROWD

EOSCpilot Science Demonstrators are pilots in the pilot, with the aim to demonstrate how disciplinary communities can make the most out of the services provided by the EOSCpilot partners. In a way, Science Demonstrators are the beta tester of the EOSC environment.

The first EOSCpilot success story is about TEXTCROWD, one of the first five funded Science Demonstrators, whose objective is to work on the collaborative semantic enrichment of text-based datasets by developing new software to enable a semantic enrichment of text sources and make it available on the EOSC.

How does it work? TEXTCROWD processes archaeological text documents, pointing out concepts about space, time and artifacts and the relations among them, in order to index the documents. It differs from the typical linguistic Natural Language Processing tools as it is based on a so-called domain ontology, i.e. a logically ordered set of concepts and categories showing their properties and the relations between them.

The goal of TEXTCROWD is to improve item searchability through the ARIADNE catalogue, as it accesses the report content in a semantic way, and not only its summary description. ARIADNE, in fact, is a catalogue for archaeological data providing access to more than 2,000,000 datasets as plain text reports of archaeological activity, that can be retrieved only through their metadata description, a sort of summary header.

The PIs were aware the success of their project was strictly related to precise domain vocabularies and on training the software. While the former was available for the chosen language (in the TEXTCROWD case, Italian), the training needed to be performed on a limited number of texts – about 100 – and required some work by experts. Nevertheless, the return is paying back this effort by large.

As a public outcome of the Science Demonstrator, TEXTCROWD will be applied on the archive (13 years of issues) of Archeologia e Calcolatori, an Italian Open Access Archaeology journal. This will enable creating an online subject index for the journal, much more effective than the usual keywords.

In terms of the EOSC being crucial in allowing the project overall fulfillment, as there is a plan to move the ARIADNE catalogue in a cloud environment, the Science Demonstrator has been the first test for moving also other related services. PIs were able to understand the steps of such process and the work required for completing it. Thus, TEXTCROWD was more critical for the EOSC as a demonstrator, than EOSC was important for TEXTCROWD as a framework, since it might have been developed as a stand-alone tool as well. Without a larger cloud framework for the data as the one we are going to create in ARIADNE, having a single service in the cloud would probably be meaningless.

TEXTCROWD might be available to the broader scientific community via other projects depending on the continuation of ARIADNE. If funding is secured for this, then it will be straightforward to port the Science Demonstrator to English as well as to other languages for which archaeological vocabularies and appropriate syntactic libraries are available. Altogether, the different language versions when coupled with the ARIADNE integrated registry will form a formidable tool for improving archaeological text findability, access and re-use. It could also be used in other, completely different domains where appropriate vocabularies are available, and the analysis of the semantic structure of the scientific discourse is well advanced through a domain ontology, as it happens for archaeology and cultural heritage.

Read more on TEXTCROWD: here's a nice poster presented at the EOSCpilot first stakeholder forum, and the Virtual Research Environment (VRE) to test its functionalities.