Scrambling for Metadata
Using Topic Modeling and Word2Vec to explore the Archives of the European Commission

Seth van Hooland,* Mathias Coeckelbergs, Ettore Rizza and Simon Hengchen

Université libre de Bruxelles (ULB)
ReSIC Research Center
Information and Communication Science Department
Avenue F.D. Roosevelt 50 – CP 123 – B-1050 Brussels, Belgium
{svhoolan,mcoeckel,erizza,shengche}@ulb.ac.be

Abstract

When and how did environmental considerations start to influence the agricultural policy development from the European Commission (EC)? What are the key documents to analyse the debate in regards to the excess production of milk and subventions to farmers? These are two examples of typical research questions historians might have in mind to address the archival fond of the EC known as the *COM collection*, consisting of documents produced by the Registry of the Secretariat-General (SG) of the European Commission spanning a period ranging from 1958 to 1982. This paper proposes an experiment with the usage of Topic Modelling (TM) and Word2Vec in order to establish an automated mapping between archival documents with descriptors of the EUROVOC thesaurus.

*Corresponding author