A New Perspective on Context - Topic-Based Information Retrieval in Archival Description*

Kerstin Arnold[1][0000-0002-4344-3798], Marta Musso[2][0000-0002-3728-3548] and Federico Nanni[0000-0002-3728-3548]

Abstract. Archives Portal Europe (www.archivesportaleurope.net) is a comprehensive and open resource on archives from and about Europe, that currently holds archival descriptions from more than 30 countries and in more than 20 languages. Following traditional approaches of archival description, the portal allows users to access the documents via the contextual entities of the records creators and the holding repositories, next to a general keyword search.

To evaluate options for subject- or topic-based access points, Archives Portal Europe is working on an automated cross-lingual topic detection tool that aims at enabling users to identify relevant documents related to a topic well beyond the narrowness of direct keyword matching. Synergising different approaches for concept-based and entity-based topics such as Natural Language Processing (NLP), Fast-Text word-embeddings aligned in a common cross-lingual "semantic" space, and retrieving name variations for entities in all kinds of languages through connections to international Linked Open Data (LOD) vocabularies such as - currently - Wikidata and the Virtual International Authority File (VIAF), the tool thereby adds to the Archives Portal Europe's existing search options.

Building on this extended discovery function, the tool, which is in its alpha development phase at the time of writing, also is meant to allow for active topic tagging in the longer term in order to improve coverage of topic-based relations between the heterogeneous and multilingual documents present in Archives Portal Europe.

This paper presents the current status quo in the portal, looks at existing options of subject-based tagging in the established approaches of archival description, in existing archival software and in current as well as emerging archival standards. It then provides an overview of the tool's set-up and summarises the initial results from the proof-of-concept phase. Last, the paper gives an outlook on next steps envisaged following the alpha and during the beta development of the tool, which will be made available as Open Source to also be of benefit for other, similar initiatives in the cultural heritage sector (see all project details on GitHub at https://github.com/ArchivesPortalEuropeFoundation/Topic-Detection/).

Keywords: Automatic Topic Detection, Archives, Information Retrieval, Vocabularies, Linked Open Data