Integrating archival materials for the study of the turbulent Greek 40s

Vicky Dritsou*†[1,2], Maria Ilvanidou*[1], Isidora Despotidou[2], Vicky Liakopoulou[2], Karmen Vourvachaki[2], and Panos Constantopoulos[1,2]

[1]Digital Curation Unit, IMSI, Athena Research and Innovation Center (DCU) – Greece
[2]Department of Informatics, Athens University of Economics and Business (AUEB) – Greece

Abstract

Humanities researchers often need to study heterogeneous digitized archives from different sources. But how can they deal with this heterogeneity, both in terms of structure and semantics? What are the digital tools they can use in order to integrate resources and study them as a whole? And what if they are unfamiliar with the methods and tools available? Towards this end, DARIAH-EU[1] and CLARIN[2] research infrastructures already support researchers in exploiting digital tools. Specific use case research scenarios have also been developed, with the PARTHENOS SSK[3] being a successful example. In this paper we describe our related (ongoing) experience from the development of the Greek research infrastructure APOLLONIS[4], where, among others, we have focused on identifying and supporting the workflows that researchers need to follow to perform specific research studies while jointly accessing disparate archives. Using the decade of 1940s as a use case, a turbulent period in Greek history due to its significant events (WWII, Occupation, Opposition, Liberation, Civil War), we have assembled (digitized) historical archives, coming from different providers and shedding light on different historical aspects of these events. From the acquisition of the resources to the desired outcome, we record the workflows of the whole research study, including the initial curation process of the digitized archives, the ingestion, the joint indexing of the data, the generation of semantic graph representations and, finally, their publication and searching. After the acquisition of the heterogeneous source materials we perform a detailed investigation of their structure and contents, in order to map the different archive metadata onto a common metadata schema, thus enabling joint indexing and establishing semantic relations among the contents of the archives. The next step is data cleaning, where messy records are cleaned and normalized. Natural Language Processing methods are then exploited for the extraction of additional information contained in the archival records or in free text metadata fields, such as persons, places, armed units, dates and topics, which enhance the initial datasets. The outcome is encoded in XML using the common schema and ingested into a repository through an aggregator implemented using the MoRE[5] system. A joint index based on a set of basic criteria is generated and maintained, thus ensuring joint access to all archival records regardless of their source. In addition, an RDF representation is generated from the encoded archival data, enabling their publication in the form of a semantic graph and supporting interesting complex queries. This is based on a specifically designed extension of CIDOC CRM[6] and a compilation of a list of research queries of varying complexity encoded in SPARQL. Preliminary tests of the entire workflows

*Speaker
†Corresponding author: v.dritsou@dcu.gr

and the tools used in all steps yielded very encouraging results. Our immediate plans include full scale ingestion and indexing of the material from a number of archives, producing the corresponding semantic graph and streamlining the incorporation of new archives.

DARIAH-EU, https://www.dariah.eu/

CLARIN, https://www.clarin.eu/

PARTHENOS Standardization Survival Kit (SSK), http://www.parthenos-project.eu/portal/ssk-2

APOLLONIS Greek Infrastructure for Digital Arts, Humanities and Language Research and Innovation, https://apollonis-infrastructure.gr/

MoRE Aggregator, http://more.dcu.gr/

CIDOC CRM, http://www.cidoc-crm.org/