Digital Muqtabas CTS Integration in CLARIN

Till Grallert
Orient-Institut Beirut
Beirut, Lebanon
till.grallert@fu-berlin.de

Jochen Tiepmar, Thomas Eckart, Dirk Goldhahn, Christoph Kuras
Natural Language Processing Group
University of Leipzig, Germany
jtiepmar@informatik.uni-leipzig.de

Abstract

This paper describes the CLARIN integration of the Canonical Text Services for a text corpus containing Muhammad Kurd Alī's *al-Muqtabas*. This high-quality text resource was introduced into CLARIN's infrastructure by using a fine grained persistently citable approach. Additionally to the practical benefits of a newly integrated text resource, this paper illustrates the usefulness of CTS as a generic interface by showing that established workflows can be reused with new data sets.

1 Introduction

In (Tiepmar et al., 2017), we introduced a newly established workflow to integrate instances of Canonical Text Services (CTS) into CLARIN's research infrastructure. In the process of extending the CTS data stock, more text collections were imported and published, including the CTS instance *muqtabas*[1], that is based on the Arabic text corpus *Digital Muqtabas*. This corpus is briefly described in this paper and in more detail in (Grallert, 2016). By incorporating new data sets like *Digital Muqtabas* in its established research infrastructure, CLARIN can provide this valuable resource to a broader audience and enhance its visibility in new user groups.

2 Digital Muqtabas

Digital Muqtabas comprises the TEI edition of all 96 issues of Muhammad Kurd Ali's monthly journal *al-Muqtabas* (The Digest / Acquired Learning) between 1906 and 1917/18 totalling some 3.8 million words. Kurd Ali (1876-1953) was one of the most influential journalists and intellectuals in Damascus (Grallert, 2016) and *al-Muqtabas* quickly became "the boldest, most coherent, consistent and committed proponent of reform and modernity (...) prior to World War I" (Seikaly, 1981). *al-Muqtabas* survived censorship and prosecution of journalists and continued publication until the final days of the war. After the war and the disintegration of the Ottoman Empire, Kurd Ali abandoned his monthly *al-Muqtabas* and turned to cultural and educational politics - ultimately serving as Syrian Minister of Education twice (Avalon, 1995).

2.1 The Source Texts

Early Arabic periodicals from the late nineteenth and early twentieth centuries, among them *al-Muqtabas* are at the core of formative discourses that still reverberate through the Arabic-speaking Middle East: the Arabic renaissance (*al-nahda*), Arab nationalism, and the Islamic reform movement. The better known and widely popular journals do not face the ultimate danger of their last copy being destroyed in the onslaught from iconoclasts, institutional neglect, and wars raging through Syria, Lybia, Yemen, and Iraq. Yet, copies are scattered throughout libraries worldwide. This makes it almost impossible to trace discourses across journals and with the demolition and closure of libraries in the Middle East, they are increasingly accessible to the affluent Western researcher only.

[1] http://cts.informatik.uni-leipzig.de/muqtabas/cts/

Among its peers, *al-Muqtabas* is one of the more readily available early Arabic periodicals and its importance to regional audiences as well as the scholarly community is underlined by a 1992 facsimile reprint. But despite this reprint and an original print run of c. 1000 copies per issue (Avalon, 1995), we were able to trace only about 40 copies globally through library catalogues and inquiries[2]. Consequently, systematic analysis of *al-Muqtabas* is all but absent from scholarly literature.

Due to the state of Arabic OCR and the particular difficulties of low-quality fonts, inks, and paper employed at the turn of the twentieth century, these texts can currently only be reliably digitised by human transcription (Märger, 2012). Funds for transcribing the tens to hundreds of thousands of pages of an average mundane periodical are simply not available. Consequently, we still have not a single digital scholarly edition of any of these journals. On the other hand, gray online-libraries of Arabic literature[3] provide access to a vast body of (mostly classical) Arabic texts including transcriptions of unknown provenance, editorial principals, and quality for some of the mentioned periodicals.

2.2 The Digital Edition

The text of our digital edition of *al-Muqtabas* was transcribed from either digital facsimiles or original copies by anonymous transcribers and uploaded to *shamela.ws*. Comparison with the original print edition allows us to discern common transcription errors. The transcription has a number of substantial and unmarked gaps, ranging from a single paragraph to multiple pages[4][5]. In terms of structural mark-up, *shamela.ws* did not provide but an eclectic selection of top-level headers. Bibliographic metadata is all but non-existing, since *shamela.ws* supplied faulty Gregorian publication dates and its own consecutive issue count that ignores the original collation into volumes.

We semi-automatically transformed the text to XML following a custom TEI schema[6] to model Arabic periodicals. At the bare minimum, we provide structural mark-up of sections and items (articles) with headings and bylines where applicable. Detailed bibliographic metadata on the issue level is provided in the TEI header.

2.3 Statistics

The TEI edition comprises all 96 issues of *al-Muqtabas* with a total of some 7'000 pages and more than 3,8 million words. In total *al-Muqtabas* printed slightly more than 5'000 articles, 3'400 of which were shorter articles in sections. In total we can currently identify only 136 different named authors[7].

3 CLARIN Integration

The text resources of the Digital Muqtabas are integrated into the CLARIN-infrastructure as described in (Tiepmar et al., 2017). For a thorough integration the granularity of text resources contained in CTS instances has to be exposed mainly via metadata. For the concrete realization the CMD profile *OLAC-DcmiTerms*[8] was used. The REST-based design of the CTS protocol reduce the effort that is necessary to include CTS instances in the center-based CLARIN infrastructure. The CLARIN center Leipzig makes strong use of webservices for both its internal

[2]Most of these are either incomplete collections or microfiche copies from one of two master fiches (produced in Chicago and Zurich).

[3]Namely *al-Maktaba al-Shāmila*, *Mishkāt*, *Ṣaydal-Fawā' id* or *al-Waraq*.

[4]See no. 5(7), pp. 463-466 and no. 5(12), pp. 761-765 for examples.

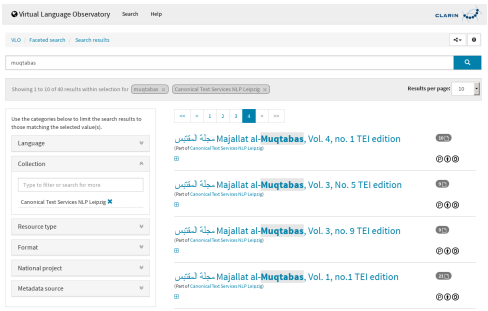
[5]Some of these could be due to pages missing from the original from which the transcribers worked or page-turning errors. Others remain enigmatic and might have been driven by editorial choices.

[6]openarabicpe_schema

[7]Two of the three most prolific among them, Ma'ruf al-Rusafi (24 articles) and Satsuna (13 articles), wrote from Baghdad. The majority of the remaining authors contributed from the cities of Greater Syria and Egypt, but some wrote from cities in France and the US.

[8]*clarin.eu* : *cr*1 : *p*_1288172614026

(a) Digital Muqtabas Text Excerpt



(b) Metadata records in the VLO (test version)

Figure 1: Digital Muqtabas and its representation in the VLO

structure and the external interfaces it provides. For the incorporation of potentially unlimited numbers of CTS instances this approach was extended by creating a wrapper webservice as main interface for the internal center infrastructure.

Regarding all metadata-centric external views the default repository system of the CLARIN center Leipzig is still used and provides a transparent interface to the CTS resources by standard interfaces like OAI-PMH. Hence, all resource can be easily found in applications like the VLO (see Figure 1b).

The implemented solution allows the creation of CMD-compliant metadata on every potential level of granularity that is provided by the CTS instance. For the time being it was decided to only expose the top two levels via metadata. For the example depicted in section 1.2 that means that the collection as a whole and every specific issue is described by metadata and can be accessed and searched for in search engines. This includes the typical descriptive metadata, all relevant references to resource-specific CTS services and the hierarchical interlinkage of metadata files as it is supported by the Virtual Language Observatory (Goosen and Eckart, 2014). Additionally, references to the content of every single article are provided.

4 Further work

The future focus of development lies on providing interfaces to other central CLARIN components. Currently an endpoint compliant to the CLARIN Federated Content Search (FCS) 2.0 is being developed and will be in productive use soon.

Acknowledgements

Part of this work was funded by the German Federal Ministry of Education and Research within the project ScaDS Dresden/Leipzig (BMBF 01IS14014B) and CLARIN-D (BMBF 01UG1120C).

References

[Avalon1995] Ami Avalon. 1995. The Press in the Arab Middle East: A History.

[Goosen and Eckart2014] Twan Goosen and Thomas Eckart. 2014. Virtual Language Observatory 3.0: What's New?. CLARIN annual conference 2014 in Soesterberg, The Netherlands.

[Grallert2016] Till Grallert. 2016. Digital Muqtabas: An open, collaborative, and scholarly digital edition of Muhammad Kurd Alī's early Arabic periodical Majallat al-Muqtabas (1906–1917/18). https://github.com/tillgrallert/digital-muqtabas.

[Märger2012] Volker Märger. 2012. Haikal Abed El: Guide to OCR for Arabic Scripts. Springer.

[Seikaly1981] Samir Seikaly. 1981. Unequal fortunes: The Arabs of Palestine and the Jews during World War I. Studia Arabica et Islamica, Beirut.

[Smith2009] David Neel Smith. 2009. Citation in classical studies. Digital Humanities Quarterly, 3.

[Tiepmar et al.2017] Jochen Tiepmar, Thomas Eckart, Dirk Goldhahn and Christoph Kuras. 2017. Integrating Canonical Text Services into CLARIN's Search Infrastructure. In Linguistics and Literature Studies.

[Tiepmar and Heyer2017] Jochen Tiepmar and Gerhard Heyer. 2017. An Overview of Canonical Text Services, Linguistics and Literature Studies.