The Semantic Web gets around to Turkic philology*

László Károly[1][0000−0001−7723−1171]

Department of Linguistics and Philology, Uppsala University, Box 635, SE-75126 Uppsala, Sweden
Laszlo.Karoly@lingfil.uu.se

Abstract. This paper presents technical aspects and considerations relevant for the creation of two novel corpus-based databases within the field of Turkic studies. The databases will provide textual and linguistic data of various Old and Middle Turkic literary languages utilising the Semantic Web technologies.
The paper evaluates the existing ontologies that the projects will implement, and provides a list of potential ontologies that philological projects are most commonly in need of.

Keywords: Turkic languages · Old Turkic · Middle Turkic · Karaim · Semantic Web · Ontology.

1 Introduction

Turkic languages form a comprehensive language family and are spoken across a wide geographical area stretching from northeast Siberia through Central Asia to the Balkans. Turkic has a long-standing written tradition beginning as early as the 8th century CE with the runiform inscriptions. Later on, various other written literary traditions emerged among the Turkic-speaking peoples when they came into contact with sedentary civilisations and religious communities.

The Turkic historical literary languages are traditionally periodised as Old Turkic, from the 8th to the 13th century, and Middle Turkic, from 13th to the beginning of the 20th century. Written sources belonging to these epochs are the subject of philological investigations.

Scholars have made a great effort to critically edit and publish the most important written documents of the Turkic languages during the last hundred years of Turkic philology. This activity has yielded several thousand invaluable publications. When it comes to digital philology, however, the situation is far from being the same. The first undertaking in this direction was the development of the VATEC database [1] in the 1990s, which now contains a small set of Old Uigur documents. Although the VATEC database provides a search engine,

* The research and writing of this paper has been supported by the European Research Council under the European Union's Horizon 2020 research and innovation programme.

its system reflects the technologies of the 1990s. During the last decades some further databases have appeared, including other materials of the Old Turkic period; see, e.g., Türik Bitig [2] and Runica [3]. They do not, however, provide real dynamic (digital) critical editions, but rather online presentations of textual materials without any use of today's technologies. To the extent of my knowledge, Middle Turkic completely lacks any form of digital or online representation.

Given the almost complete lack of modern, technologically advanced databases for the Old and Middle Turkic literary languages, two new initiatives have been started to address these gaps. One project aims at creating a new database for the Turkic runiform inscriptions. On the one hand, this is a philological project to reedit the already available materials, and on the other hand, it is addressing the challenge of establishing a database that makes use of the technologies of the Semantic Web. The other project is more ambitious: it aims to create the first scholarly database of Middle Turkic, also utilising, among other things, Semantic Web technologies.

Since the two projects differ greatly in many respects, they will be described separately in the following sections.

2 The "Uppsala" database of Turkic runiform inscriptions

The Database of Turkic Runiform Inscriptions was established at the University of Mainz, Germany, in the course of 2015 as a start-up project financed from university sources [4]. It is now hosted by Uppsala University at http://www.runiform.lingfil.uu.se/ [5].

At the time of its creation, the team decided to use MediaWiki (MW) [6] and Semantic MediaWiki (SMW) [7] as major components of the underlying system. Although MW is primarily designed for Wikipedia, and thus has its own internal limitations, the combination of MW & SMW provides a flexible and easily configurable software environment with an extent set of semantic web features ready for immediate use. For instance, complex browsing of data is facilitated by the Semantic Compound Queries extension, showing data on maps is allowed by the Semantic Maps extension, and Semantic Result Formats present query results in different formats such as calendars, timelines, charts, filterable results or graphs.[1]

For the time being, the database implements some thirty individual properties in order to store semantic information. The property declarations and names are, however, preliminary. They do not correspond to any element of a standardised ontology, but were merely created for testing purposes. On ontologies for our philological/linguistic needs, see below. The semantic information already stored in individual pages (descriptions of runiform inscriptions) makes it possible to automatically collect and then present various pieces of information.

[1] For the time being our SMW stores semantic information in the same relational database that is used by the whole platform. Separating semantic data in an RDF store is open for future consideration. After data separation, SMW could support the OWL ontology language and offer SPARQL web services for querying the datasets.

3 A database of Middle Turkic written documents

Middle Turkic is an epoch including tens of thousands of manuscripts from different parts of Eurasia in several different writing systems; therefore the creation of such a voluminous database requires careful planning. Starting on a small-scale basis, our team selected a relatively small, but internally homogeneous set of sources: the Karaim Bible translations in Hebrew script. This group of manuscripts is suitable for the establishment of the database and for testing all the necessary software components of the platform.

The implementation will be done within the framework of the ERC-funded starting grant (*Re*)*constructing a Bible. A new approach to unedited Biblical manuscripts as sources for the early history of the Karaim language* (*KaraimBible*) headed by Michał Németh, Jagellonian University, Poland, project period: 1 February 2019–31 January 2024 [8].

The platform will have a multi-layered organisation including annotation layers from 0 to 3. A complex network of ontologies will be responsible for layer-internal relations and dependencies, on the one hand, and for interrelations between the individual layers, on the other.

Layer 0 (*Diplomatic editions*): this is the layer for storing individual manuscripts as transcribed texts. The properties of a manuscript will be stored in a form compatible with the *Text Encoding Initiative* (TEI) data modelling [9], but with extended semantics. As part of the semantic annotation, words of the transcribed texts will be tagged automatically by using word-tag and suffix-tag associations. The output of Layer 0 will present the manuscripts as regular diplomatic editions with various commentaries related to their philological or linguistic peculiarities. Given the richly annotated texts, glossaries and various lexicographical data concerning individual manuscripts can also be visualised in Layer 0.

Layer 1 (*Critical editions*): this is the layer for storing manuscript information related to different copies of the same text. Layer 1 will allow sentence-based parallel representation which can help scholars to identify differences and changes occurring over time and space, which are marked automatically in colour. Another function of Layer 1 will be to present related manuscripts in the form of classical critical editions.

Layer 2 (*Literary languages*): this layer will join the sources into groups covering specific areas and/or eras, for example, combining South- and North-Western Karaim sources into a Western Karaim group. Language-specific glossaries or lexicons will also be provided in Layer 2 by collecting and uniting data stored in Layer 0. A sub-layer will be added to represent and handle dialectal differences. This layer will also provide language-specific information and allow for making comparisons between Middle Turkic languages.

Layer 3 (*Middle Turkic*): this level will handle Middle Turkic literary languages as a whole, and will thus be responsible for overarching representation. The complete lexicon of Middle Turkic will be generated on this level.

The project has only recently begun, and no final decision has been made on which software components would best meet our needs and requirements.

There are several possible alternatives that vary significantly when it comes to adaptability and sustainability. The Apache Jena framework [10] seems to be a strong candidate for the project, providing the necessary components, e.g., GRDDL to transform XML to RDF. Another alternative is the XTriples infrastructure [11,12], developed at the *Academy of Sciences and Literature, Mainz*. In the case of the latter, it would be straightforward to use eXist-db [13] and TEI Publisher [14] for publishing content on the web.

4 The ontologies: needs and reality

In addition to the software components needed for building Semantic Web and Linked Data applications, the underlying ontology (or more precisely a set of intertwining ontologies) plays a significant role in organising the information into useful data and knowledge. Established and commonly accepted ontologies exist for commercial purposes. Medicine and branches of the natural sciences have also made use of them in several forms. The humanities, however, lag behind in their adaptation and application.

The discussion has already started; for instance, theoretical considerations have been presented [15], and promising trials have been undertaken to furnish the widely used TEI encoding scheme, or at least some of its subsets, with semantics [16]. An example of practical application is the Sharing Ancient Wisdoms (SAWS) project [17,18] in the humanities community. The SAWS team (1) designed an ontology for the purpose of recording and visualising links within and between textual materials, and (2) created RDF data extracted from the given textual material encoded in TEI/XML. The project used a subset of TEI called "TEI for Gnomologia" [19] based on TEI for Manuscripts.

Given our two projects dealing with textual materials from various Turkic languages, the task of developing a TEI-compatible ontology needs immediate attention. Due to the complexity of TEI, it is reasonable, as others have done, to create a customisation, that is, to define a set of module and attribute classes that meet our requirements. As a specific subset of TEI selected for the description of epigraphic materials, TEI-EpiDoc [20] seems to be a reasonable starting point. If our needs go beyond what EpiDoc offers, further TEI elements can easily be added.

Although preliminary work has been done toward developing an ontology for handling epigraphic information, or more specifically TEI-EpiDoc [21,22], no single model is currently available for use. Two existing ontologies can however be immediately utilised in our projects: (1) the SAWS ontology, which provides a vocabulary for the TEI subset "Gnomologia" with an overlap with TEI-EpiDoc, and (2) the LAWD ontology designed for Linked Ancient World Data [23].

An ontology for describing bibliographic resources will be a key component of our runiform database. It will store information about more than 3000 scientific publications in the field of runiform studies. For this purpose, we will import the FaBiO vocabulary [24] into our MW & SMW database. The implementation of FaBiO in the database of Middle Turkic sources is a future option.

Despite the strong philological orientation of the projects, linguistic investigations of the textual materials will also be carried out. Linguistic information is planned to be provided through the LexInfo ontology [25].[2] Finally, semantic data related to the various writing systems will also be modelled. A plausible proposal for this is the ontology for accessing transcription systems [27].

In accordance with the above-stated objectives, the following ontologies are planned to be implemented in our databases:

1. Textual/epigraphic information: SAWS, LAWD, OEDUc/EPIDOC or equivalent. Significant gaps, far from ready for use.
2. Bibliographical information: FaBiO. Complete, ready to implement.
3. Linguistic information: LexInfo/OLiA. Complete, ready to implement.
4. Writing system-related information: OATS. A proposal, far from ready for use.

5 Conclusion

As one of the academic disciplines with limited human resources, Turkic studies lags behind significantly, even within the humanities, in the creation of databases with Semantic Web technologies. We hope that our initiatives can fill some of the obvious gaps, both within our respective field of study and within Digital Humanities as a whole, and, at the same time, provide knowledge and information for the broader community in the form of Linked Open Data.

References

1. VATEC = Erdal, M. & Gippert, J. & Röhrborn, K. & Zieme, P. (eds.). Vorislamische Alttürkische Texte: Elektronisches Corpus (1999–2003) http://vatec2.fkidg1.uni-frankfurt.de/. Last accessed 7 Feb 2019
2. Türik Bitig, http://bitig.org. Last accessed 11 Feb 2019
3. Runica, http://www.altay.uni-frankfurt.de/english/runika_eng.htm. Last accessed 11 Feb 2019
4. Károly, L. & Rentzsch, J.: An online database of Turkic runiform inscriptions. In: Telicin, N. N. & Šen, J. N. (eds.) Actual problems of Turkic Studies, 534–541. Sankt-Peterburg: SPbGU Kafedra tjurkskoj filologii (2016). http://orient.spbu.ru/books/actual_problems_turkic_studies_180/534/
5. Károly, L. & Rentzsch, J. (eds.): A database of Turkic runiform inscriptions (2015–2019) http://www.runiform.lingfil.uu.se/. Last accessed 7 Feb 2019
6. MediaWiki, https://www.mediawiki.org. Last accessed 25 Feb 2019
7. Semantic MediaWiki, https://www.semantic-mediawiki.org. Last accessed 25 Feb 2019
8. ERC Starting grant 2018 examples, https://erc.europa.eu/news/erc-starting-grant-2018-examples. Last accessed 11 Feb 2019

[2] The applicability of the Ontologies of Linguistic Annotation (OLiA) framework [26] will also be evaluated.

9. TEI Consortium: TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 3.5.0. Last updated on 29 January 2019, revision 3c0c64ec4 (2019). https://jenkins.tei-c.org/job/TEIP5/lastSuccessfulBuild/artifact/P5/release/doc/tei-p5-doc/en/html/index.html
10. Apache Jena, https://jena.apache.org. Last accessed 13 Feb 2019
11. XTriples. A generic webservice to extract RDF statements from XML resources, http://xtriples.spatialhumanities.de. Last accessed 17 Feb 2019
12. Grüntgens, M. & Schrade, T.: Data repositories in the Humanities and the Semantic Web: modelling, linking, visualising. In: Proceedings of the 1st Workshop on Humanities in the Semantic Web co-located with 13th ESWC Conference 2016 (ESWC 2016), Anissaras, Greece, May 29th, 2016, 53–64. (2016). http://ceur-ws.org/Vol-1608/paper-07.pdf. Last accessed 17 Feb 2019
13. EXist-db. The Open Source Native XML Database, http://exist-db.org. Last accessed 17 Feb 2019
14. TEI Publisher. The Instant Publishing Toolbox, https://teipublisher.com. Last accessed 17 Feb 2019
15. Eide, Ø.: Ontologies, Data Modeling, and TEI. Journal of the Text Encoding Initiative **8**, 1–22 (2015). http://jtei.revues.org/1191.doi:10.4000/jtei.1191
16. Ciotti, F. & Tomasi, F.: Formal Ontologies, Linked Data, and TEI Semantics. Journal of the Text Encoding Initiative **9**, 1–22 (2016). https://doi.org/10.4000/jtei.1480
17. Jordanous, A. & Roueché, Ch. & Tupman, Ch. & Lawrence, K. F. & Hedges, M. & Wakelnig, E. & Searby, D.: Sharing Ancient Wisdoms (SAWS) ontology (2013) http://www.ancientwisdoms.ac.uk/. Last accessed 15 Feb 2019
18. Tupman, Ch. & Jordanous, A.: Sharing ancient wisdoms across the Semantic Web using TEI and ontologies. In Andrews, T. & Macé, C. (eds.) Analysis of Ancient and Medieval Texts and Manuscripts: Digital Approaches, 213–228. Turnhout: Brepols (2014). https://doi.org/10.1484/M.LECTIO-EB.5.102572
19. TEI for Gnomologia, http://www.ancientwisdoms.ac.uk/media/documents/Markup_Guidelines_for_Gnomologia.html. Last accessed 17 Feb 2019
20. Elliott, E. & Bodard, G. & Cayless, H. *et al.*: EpiDoc: Epigraphic Documents in TEI XML. Online material (2006–2017) http://epidoc.sf.net. Last accessed 17 Feb 2019
21. Álvarez F. L. & García-Barriocanal E. & Gómez-Pantoja J. L.: Sharing Epigraphic Information as Linked Data. In: Sánchez-Alonso S., Athanasiadis I.N. (eds) Metadata and Semantic Research. MTSR 2010, Communications in Computer and Information Science, Vol. 108, 222–234. Berlin & Heidelberg: Springer (2010). https://doi.org/10.1007/978-3-642-16552-8_21
22. EpiDoc/OEDUc. Epigraphic ontology, https://github.com/EpiDoc/OEDUc/wiki/Epigraphic-ontology. Last accessed 17 Feb 2019.
23. Cayless, H. A.: LAWD. Linking Ancient World Data Ontology (2013) https://github.com/lawdi/LAWD. Last accessed 17 Feb 2019
24. Peroni, S. & Shotton, D.: FaBiO and CiTO: Ontologies for describing bibliographic resources and citations. Web Semantics: Science, Services and Agents on the World Wide Web **17**, 33–43 (2012). https://doi.org/10.1016/j.websem.2012.08.001
25. LexInfo, https://lexinfo.net. Last accessed 17 Feb 2019
26. Chiarcos, Ch. & Sukhareva, M.: OLiA – Ontologies of Linguistic Annotation, Semantic Web Journal **6(4)**, 379–386 (2015). http://semantic-web-journal.net/system/files/swj518_0.pdf
27. Moran, S.: An ontology for accessing transcription systems. Lang Resources & Evaluation **45**, 345–360 (2011). https://doi.org/10.1007/s10579-011-9158-8