JADH 2018

“Leveraging Open Data”

September 9-11, 2018
Hitotsubashi Hall, Tokyo

https://conf2018.jadh.org

Proceedings of the 8th Conference of Japanese Association for Digital Humanities

ALLIANCE OF
DIGITAL
HUMANITIES
ORGANIZATIONS

TEI
TEXT ENCODING INITIATIVE

TEI 2018

“TEI as a Global Language”

September 9-13, 2018
Hitotsubashi Hall, Tokyo

https://tei2018.dhii.asia

Book of Abstracts
The 18th Annual TEI Conference and Members’ Meeting

In the field of classical Japanese literature, researchers and students conduct research based on ancient text which often has textual variations. In such a case, research starts with comparing those texts and manually listing their differences. However, the author found out the fact that by applying TEI structure to the text, one can leave the manual task to the machine. Moreover, he also found out that by using well-known literary works, such as the Kokin-wakashu (the oldest imperial anthology of Japanese poems), one can help students' better understanding about how the TEI could be involved with classic Japanese literary works, which eventually develops their interests in both TEI and the literary works.

TEI Lex-0: A Target Format for TEI-Encoded Dictionaries and Lexical Resources

Laurent Romary[1], Toma Tasovac[2]

Achieving consistent encoding within a given community of practice has been a recurrent issue for the TEI Guidelines. The topic is of particular importance for lexical data if we think of the potential wealth of content we could gain from pooling together the information available in the variety of highly structured, historical and contemporary lexical resources. Still, the encoding possibilities offered by the Dictionaries Chapter in the Guidelines are too numerous and too flexible to guarantee sufficient interoperability and a coherent model for searching, visualising or enriching multiple lexical resources.

Following the spirit of TEI Analytics [Zillig, 2009], developed in the context of the MONK project, TEI Lex-0 aims at establishing a target format to facilitate the interoperability of heterogeneously encoded lexical resources. This is important both in the context of building lexical infrastructures as such [Ermolaev and Tasovac, 2012] and in the context of developing generic TEI-aware tools such as dictionary viewers and profilers. The format itself should not necessarily be one which is used for editing or managing individual resources, but one to which they can be univocally transformed to be queried, visualised, or mined in a uniform way. We are also aiming to stay as aligned as possible with the TEI subset developed in conjunction with the revision of the ISO LMF (Lexical Markup Framework) standard so that coherent design guidelines can be provided to the community (cf. [Romary, 2015]).

The paper will provide an overview of the various domains covered by TEI Lex-0 and the main decisions that were taken over the last 18 months: constraining the general structure of a lexical entry; offering mechanisms to overcome the limits of <entry> when used in retro-digitized dictionaries (by allowing, for instance, <pc> and <lbl> as children of <entry>); systematizing the representation of morpho-syntactic information [Bański et al., 2017]; providing a strict <sense>-based encoding of sense-related information; deprecating <hom>; dealing with internal and external references in dictionary entries, providing more advanced encodings of etymology (see submission by Bowers, Herold and Romary); as well as defining technical constraints on the systematic use of @xml:id at different levels of the dictionary microstructure. The activity of the group has already lead to changes in the Guidelines in response to specific GitHub tickets[3].

Acknowledgements

The initiative started in the context of the COST Action European Network for e-Lexicography (ENEL) and is now supported by DARIAH, as part of the activities of the

[1] Inria (team ALMAnaCH), France and DARIAH

[2] Belgrade Center for Digital Humanities, Serbia

[3] See for instance https://github.com/TEIC/TEI/issues/1702 (model.entryPart.top for <pc> and <lbl>), https://github.com/TEIC/TEI/issues/1688 (add <form> to att.typed) or https://github.com/TEIC/TEI/issues/1734 (make hyph/stress/syll members of att.notated)

Lexical Resources working group[1], as well as the H2020-funded project European Lexicographic Infrastructure (ELEXIS)[2].

References

Bański, Piotr, Jack Bowers and Tomaz Erjavec (2017). TEI-Lex0 guidelines for the encoding of dictionary information on written and spoken forms. *Electronic Lexicography in the 21st Century: Proceedings of ELex 2017 Conference*, Sep Leiden, Netherlands. 〈hal-01757108〉

Ermolaev, Natalia, and Toma Tasovac (2012) “Building a Lexicographic Infrastructure for Serbian Digital Libraries.” *Libraries in the Digital Age (LIDA) Proceedings* 12, no. 0 (June 12). http://ozk.unizd.hr/proceedings/index.php/lida/article/view/55.

ISO 24613:2008 *Language resource management - Lexical markup framework (LMF)*, currently revised as a multipart standards with Part 1: core model, Part 2: Machine Readable Dictionaries, Part 3: Etymology, Part 4: TEI serialisation

Romary, Laurent (2015). TEI and LMF crosswalks. *JLCL - Journal for Language Technology and Computational Linguistics,* 30 (1), 〈http://www.jlcl.org〉. 〈hal-00762664v4〉

Zillig, Brian (2009) “TEI Analytics: Converting Documents into a TEI Format for Cross-Collection Text Analysis.” *Literary and Linguistic Computing* 24 (2009): 187–192. https://doi.org/10.1093/llc/fqp005

[1] https://www.dariah.eu/activities/working-groups/lexical-resources/

[2] https://www.cordis.europa.eu/project/rcn/213379_en.html