The Journal of Open Source Software

teiphy: General-purpose Python utility for converting TEI XML collations to NEXUS and other formats

James McCollum [1] **and Robert Turnbull** [2]

1 Institute for Religion and Critical Inquiry, Australian Catholic University, Australia **2** Melbourne Data Analytics Platform, University of Melboune, Australia

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

Summary

Textual scholars have been using phylogenetics to analyze manuscript traditions since the early 1990s (Robinson & O'Hara, 1992). Many standard phylogenetic software packages accept as input the NEXUS file format (Maddison et al., 1997). The teiphy program takes a collation of texts encoded using the Text Encoding Initiative (TEI) guidelines and converts it to a NEXUS format so that it can be used for phylogenetic analysis. It can also convert to other formats as well.

Statement of need

The TEI is an initiative to provide an international standard for digital encoding textual information for the humanities (Ide & Sperberg-McQueen, 1995). The TEI guidelines express a standard XML format for encoding a critical apparatus (Consortium, 2022).

discussion of TEI

Design

basic history of phylogenetics and texts

other formats

Design

While this is a straightforward process for most collation data, lacunae, retroversions, and other sources of ambiguity occasionally make a one-to-one mapping of witnesses to readings impossible, and in some cases, one disambiguation may be more likely than another in a quantifiable way. Mechanisms for accommodating such situations exist in both TEI XML and NEXUS, and for likelihood-based phylogenetic methods, "soft decisions" about the states at the leaves and even the root of the tree can provide useful information to the inference process. For these reasons, I wanted to ensure that these types of judgments, as well as other rich features from TEI XML, could be respected (and, where, necessary, preserved) in the conversion process.

Collations should preserve as much detail as possible, including information on how certain types of data can be normalized and collapsed for analysis. Since we may want to conduct the same analysis at different levels of granularity, the underlying collation data should be available for us to use in any case, and only the output should reflect changes in the desired level of detail. Likewise, as noted in the previous section, uncertainty about witnesses' attestations should be encoded in the collation and preserved in the conversion of the collation.

For text-critical purposes, differences in granularity typically concern which types of variant readings we consider important for analysis. At the lowest level, readings with uncertain or reconstructed portions are almost always considered identical with their reconstructions for the purpose of analysis. Defective forms that are obvious misspellings of a more substantive reading are often treated the same way. Even orthographic subvariants that reflect equally "correct" regional spelling practices may be considered too common and of too trivial a nature to be of value for analysis. Other readings that do not fall under these rubrics but are nevertheless considered manifestly secondary (due to late and/or isolated attestion, for instance), may also be considered uninformative "noise" that is better left filtered out.

Use Case

Ephesians UBS example? Is there an example from

Availability

The software can be installed through the Python Package Index (PyPI) and the source code is available under the MIT license from the Github repository. The automated testing suite has 100% coverage.

Acknowledgements

Stephen Carlson.

References

Consortium, T. (2022). *TEI P5: Guidelines for electronic text encoding and interchange: Critical apparatus [v.4.4.0]*. https://www.tei-c.org/release/doc/tei-p5-doc/en/html/TC.html.

Ide, N. M., & Sperberg-McQueen, C. M. (1995). The TEI: History, Goals and Future. *Computers and the Humanities*, *29*(1), 5–15.

Maddison, D. R., Swofford, D. L., & Maddison, W. P. (1997). Nexus: An Extensible File Format for Systematic Information. *Systematic Biology*, *46*(4), 590–621. https://doi.org/10.1093/sysbio/46.4.590

Robinson, P., & O'Hara, R. J. (1992). Report on the Textual Criticism Challenge 1991. *Bryn Mawr Classical Review*, *3*(4), 331–337.