

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

James McCollum¹ and Robert Turnbull²

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

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

Software

- [Review](#)
- [Repository](#)
- [Archive](#)

Editor: [Open Journals](#)

Reviewers:

- [@openjournals](#)

Submitted: 01 January 1970

Published: unpublished

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

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

35 detail. Likewise, as noted in the previous section, uncertainty about witnesses' attestations
36 should be encoded in the collation and preserved in the conversion of the collation.

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

46 Use Case

47 Ephesians UBS example? Is there an example from

48 Availability

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

52 Acknowledgements

53 Stephen Carlson.

54 References

- 55 Consortium, T. (2022). *TEI P5: Guidelines for electronic text encoding and interchange: Critical*
56 *apparatus* [v.4.4.0]. <https://www.tei-c.org/release/doc/tei-p5-doc/en/html/TC.html>.
- 57 Ide, N. M., & Sperberg-McQueen, C. M. (1995). The TEI: History, Goals and Future.
58 *Computers and the Humanities*, 29(1), 5–15.
- 59 Maddison, D. R., Swofford, D. L., & Maddison, W. P. (1997). Nexus: An Extensible
60 File Format for Systematic Information. *Systematic Biology*, 46(4), 590–621. <https://doi.org/10.1093/sysbio/46.4.590>
- 62 Robinson, P., & O'Hara, R. J. (1992). Report on the Textual Criticism Challenge 1991. *Bryn*
63 *Mawr Classical Review*, 3(4), 331–337.