# Decoding Text and Music: TEI (and MEI) in the Thesaurus Musicarum Latinarum

The Thesaurus Musicarum Latinarum (TML) is a searchable archive of music theory treatises in Latin dating from late antiquity to the seventeenth century. Over the last two decades, the TML has established itself as one of the most valuable and authoritative online resources in the field of musicology, but it is also well known to scholars of other disciplines interested in documenting the broader intersection of arts and sciences within the Western tradition. The TML consists of approximately 1100 treatises that discuss general principles (e.g. the origin and nature of music, its role within human activities) or technical issues (e.g. the formation of scales, the construction of musical instruments, their tuning); define genres or styles (within liturgical, secular, and instrumental repertories); or prescribe rules about composition, notation and to some extent performance. Many of the treatises include illustrations and drawings exemplifying the issues and rules discussed and musical examples, sometimes derived from actual compositions (sometimes the only remnants of otherwise lost pieces). The TML is currently being encoded with TEI; at present, a majority of the 14th- and 15th-century texts are encoded.

These writings are particularly rich in references, direct or indirect, not always explicit or complete, internal or from other musical and non-musical writings. So in addition to the usual structural markup and header information, selected details of the content have been encoded and for special purposes linked to one another to capture the whole information about cited texts. For example, each mention of a person, place, and work title is marked with its appropriate tag and assigned an xml:id. These elements are then associated with any other related elements, including quotes and paraphrases of cited texts. This markup allows for not only greatly enhanced viewing capabilities but also more granular and specialized searching and browsing, and creates new ways to analyze this corpus. The encoding has also been used to create an interface for comparing multiple editions of the same text using Juxta, including encoded transcriptions of manuscripts all of which are to be linked to images of the original sources.

Enrichment and standardization of the original texts allows users to find all works (textual or musical) mentioned by a title, and any people, places, or works mentioned in each text, tasks normally difficult when dealing with an inflected language. In addition, the markup of place names allows, in conjunction with the Getty Thesaurus of Geographic Names and a geotag service, the implementation of a map-based interface for browsing the corpus by location. We will also use this information to integrate into the XTF reader the ability to browse through a text geospatially and find related information.

The encoding also allows the separation of all paratextual elements (openers, closers, section titles, figure captions and notes) from the core text, making a traditional search more precise. In fact, the texts of this corpus are focused on a relatively small number of musical subjects often inconsistently invoked throughout a treatise, but whose specific discussion is best discovered in rubrics, titles, and other paratextual elements.

On the other hand, the reverse holds true for broader computational analysis of the corpus and subsets of texts. By excluding the paratext we are able to generate more accurate topic models using latent Dirichlet allocation (LDA) in order to create lists of related texts. Through a combination of custom R

scripts (based around the topicmodels package) and the d3 javascript library we have created an interface that allows users to browse through the corpus topically. We also use the encoding to extract citations which can then be made more complete using authority files to fill in missing information. Using these citations we are able to generate citation maps of the corpus and relevant subsets using Network Workbench. Since the TEI will be freely available, researchers will be able to easily find and process texts (or portions of texts) relevant to their interests for their own analysis.

However, because music treatises contain non-textual components, in this case, over 15,000 musical examples intertwined with the texts, TEI alone is not sufficient to represent them. Adding MEI to the encoding procedures of the TML corpus offers a number of potential advantages for this project. Musical examples in the TML are currently either reproduced as graphics or partially encoded in a custom code, but features for searching, browsing and graphical display of musical notation were never developed. We have recently started transforming the old code into MEI (Music Encoding Initiative, whose Guidelines were first released in 2013) and encoding the musical examples anew. MEI is a very flexible encoding model that not only features an architecture that is meant to be easily accommodated into a TEI document but also creates the possibility to further align notation to graphics and sound. Some of the procedures adopted for encoding and decoding the verbal dimension of this corpus are suitable to be extended to its musical dimension with MEI, which will allow for an exhaustive investigation of the TML corpus, first and foremost to search for particular musical phrases, and perform analysis of frequency of certain intervals or rhythmic patterns. The notation found in the texts currently being encoded (late-medieval mensural notation) is not extensively covered by the current MEI guidelines, so this project is also an opportunity to further extend the MEI model.

In sum, this presentation will show how TEI (and MEI) encoding enables increased discoverability of texts, aids in preforming analysis of the corpus and generates new ways of visualizing the relationships between and within texts as well as allowing us to identify and analyze subsets of documents that are relevant to individual research needs. For instance, a researcher will be able to investigate the co-citations of a specific author and how this citation network changed over time, what topics were associated with documents that mention specific geographical regions, or how alterations in the musical examples used to illustrate the same text perhaps responded to changing musical taste.

Working Bibliography

Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84.

Elliott, T., & Gillies, S. (2009). Digital Geography and Classics, 3(1). Retrieved from http://www.digitalhumanities.org/dhq/vol/003/1/000031.html

Hornik, K., & Grün, B. (2011). topicmodels: An R package for fitting topic models. Journal of Statistical Software, 40(13), 1-30.

The SIMILE Project, http://simile.mit.edu/

XTF http://xtf.cdlib.org/