Enabling the findability, accessibility and reusability of corpus data.
Deep-structured model-based metadata for historical corpora
Carolin Odebrecht, Humboldt-Universität zu Berlin

Long-term storage and documentation are two requirements (among others) linked to corpus data management (Wilkinson et al. 2016). Repository systems provide services for data storage. Metadata often provide the base of corpus data documentation. Open access to corpora is crucial condition for different reuse scenarios such as study replication, application of new studies, or enrichment of existing data. Before reusing corpora becomes possible, researchers need to search for and find them. Similar to the search and visualization of annotations within the corpora, the search (and visualization) of an entire corpus data set is a challenging but necessary task. Especially when we take into account that historical corpora vary with respect to the corpus design, annotations schemes and formats. Following Bird and Simons (2001: 8), researchers may either deal with low recall – if their search reveals not all existing data – or with low precision – if the search reveals entirely irrelevant matches. How, then, can we design corpus metadata for these kind of purposes?

In my talk, I would like to discuss the findability, accessibility and reusability of corpus data and how a model-based metadata scheme can address these issues. First, metadata need to describe heterogeneous historical corpora in a uniform and extensive way (accessible to researchers). Second; metadata should provide all necessary information to enable scenarios (reusability). Third, metadata should support a search system with as high as possible recall and precision (findability). The Metamodel for Corpus Metadata (MCM) (Odebrecht 2014) can address these three functionalities independent from the respective format or research context of the historical corpora.


Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da Silva Santos, Luiz Bonino; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim et al. (2016): The FAIR Guiding Principles for scientific data management and stewardship. In: Scientific data 3. 160018. doi: 10.1038/sdata.2016.18