AOR Data Release 3.0: Status of the transcriptions

30 August 2016

We have not added any new transcriptions since the previous data release (July 2016), but we continued our process of checking existing transcriptions and a number of transcriptions have been modified (typographical errors removed, missing marginal annotations added, &c.). While preparing our data files for the data analysts, we discovered a bug that caused some incorrect word frequency calculations. Additionally, some wrongly attributed language tags caused certain vocabulary files to be polluted by the inclusion of words from another language. These errors have been fixed and our dataset has been cleaned up significantly since the AOR Data Release 2.0.

In another instance a temporary solution involving marginalia that run across multiple pages may slightly skew statistics, limited only to 41 instances.. We have created a way to link these marginal notes, or rather their constituent parts, in XML, but automatically joining them up in the transcription panel of our viewer will be a feature for development in the next phase of AOR. Since we need to make clear the existence of these "running marginalia" to our users, we have included in square brackets a reference to the page number where the remaining part of a marginal note can be found, as well as the relevant text. For example: "[p. 346: Plurimis literulis fit homo modica[e] habilitatis:] Autotechnia, et Cosmopraxia Vnica." **Because of this solution, the text in between the square brackets will be counted twice in any search**, something the users of this data release should be aware of, even though this applies only to a tiny fraction of our dataset.

Another difference with the previous data releases is that we have stripped out the square brackets in the data files that are part of the data release. Those specific brackets were used by transcribers to signify uncertainties of interpretation, but more often to denote potentially confusing ligatures and abbreviations. Those brackets ended up cutting off some words, also influencing user search statistics. As a result, we decided not to take these brackets into account when creating the files for this latest data release.

We have made some progress in dealing with the around 75 remaining questions on the transcriptions, ranging from the translation of enigmatic marginal notes to trying to decipher vague marks and unfamiliar symbols. A tiny minority of Harvey's annotations therefore remain shrouded in uncertainty; nonetheless, all of his annotations are transcribed and, more importantly, the uncertainty the transcribers might have felt about some of his annotations remain highlighted in the transcriptions themselves for users to be aware of. It remains our goal to resolve all these questions connected to the AOR Harvey corpus, but if this turns out to be impossible, the project team will upload an enumerated list of all outstanding questions to the AOR website. For more information on the dataset, please consult the user documentation accompanying the previous two data releases or, for more detailed information, the current, updated Transcriber's Manual. Both documents can be found on the downloads page of the AOR website.

The structure of the data release has undergone some changes, information about which you can find here.

Suggested citation:

Archaeology of Reading Data Release, version 3.0 (August 2016), based on data from www.bookwheel.org