# AOR Data Release 1.0: Status of the Transcriptions
## 30 November 2015

The majority of the files of which this AOR Data Release 1.0 consists constitute the transcriptions of all of Gabriel Harvey's annotations made in the twelve books that comprise the AOR Phase 1 Harvey corpus. All of Harvey's annotations are captured in XML, and the files contained in this Data Release capture the original transcriptions. Although at this point in the project all of Harvey's annotations have been transcribed, second-person cross-checking of all the transcriptions is still ongoing. There are also some remaining issues, such as difficult passages in Latin or Greek, as yet unrecognizable symbols, and unclear references to titles of books or names of people, for which we need the input of specialists. This document specifies the status of the transcriptions per each book (not per each individual transcription), in order to inform those interested in the AOR transcription dataset about the particular attributes, and current limitations, of the data pending final cross-checking.

Before addressing these attributes and limitations, it remains to describe workflow issues and to outline the process by which members of the AOR humanities content team generated and checked transcriptions. Once any initial transcription of an annotated page of one of the books within the AOR Phase 1 Harvey corpus was completed, the transcriber uploaded it to the AOR GitHub repository, assigning to it a commit message "First pass – to be checked." These transcriptions were then cross-checked by another transcriber on the content team, who finalized that relevant transcription, if everything recorded by the first transcriber was correct. When errors or variances were found, the cross-checker would alternatively insert comments in the XML specifying these mistakes, and then attach a commit message to the file in order to reflect the remaining work that still needed to be done on it. This second version of the file was then uploaded to the AOR GitHub repository, where it was preserved not only in its latest version, but also alongside all previous iterations of the file. This system of simultaneous version preservation and control made it possible to view all changes made to the files before a final step was taken. Only after any necessary final changes were made would the file will be marked as "finalised" (i.e., transcribed, checked, corrected if need be, and finished).

In general, this Data Release 1.0 presents all the initial versions of all Harvey transcriptions that have been made to date. Subsequent to this Data Release 1.0, further work is being done to cross-check transcriptions and modify some transcriptions based on recent changes made to the AOR Phase 1 XML schema. Note that these recent changes to the schema do not generate new content, per se, since all of Harvey's annotations have been captured already. Only the transcriptions which have not been cross-checked may still remain incomplete. Those transcriptions that have not yet been cross-checked may also include other mistakes, such as tagging the wrong named persons or books in a given Harvey annotations, as well as typographical errors, or imperfect translations of the non-English marginal notes into the English vernacular. These details must be taken into account by anyone who is interested to download our dataset and begin working with it in its current state of development.

Despite these, admittedly minor, limitations in our dataset, it nonetheless provides a wealth of information about Gabriel Harvey's annotations and historical reading practices, from the interpretive content of the annotations themselves, to the hundreds of people and historical figures, geographical locations, and other texts he identifies in these marginal notes. These, too, have been tagged with standardized identifiers, making it possible to search for such information both within and across the corpus. A list of the people, places, and cross-references to other books tagged can also be found in these Excel files, which also comprise a portion of this larger data release: book.csv, locations.csv, peole.csv.

Some of the data which are contained in .csv files are list with words and the frequency with which they appear in marginal annotations. These files are encoded in UTF-8 and the normal Windows encoding will result in gibberish when opening them. In order to avoid this, open Microsoft Excel, go to the tab 'Data', select 'From text', select the .csv file you want to open, press 'Import' and then select for file origin 'Unicode (UTF-8). This applies to all the .csv files of which the name starts with 'vocab_'.

The way in which the data is structured reflects the various forms of annotation used by Harvey. The four main categories, which comprise the vast majority of the dataset, are:

1) **Marginal notes** (and their translation).
2) **Symbols**
3) **Marks**
4) **Underline**

Other tags are used to capture less common forms of annotation employed by Harvey, such as drawing (Harvey's sparse attempts to draw something in his books), errata (Harvey changing the spelling of a word), and numeral (Harvey inserting numbers that are standing on their own, i.e. are not part of a marginal note). Metadata have been used to add information to the all forms of annotation, such as capturing the hand and language in which the marginal note was written, the position of a reader's intervention on the page, or the language of the printed text.

Furthermore, since the transcriptions consists of all annotations on one page, every XML file contains information about that particular page, such as pagination and/or signature, thus specifying the location of that page within the book. All of which is, in a nutshell, the most important data captured by our schema. The complete AoR schema is part of the data release but can also be accessed here. Additional information can also be found in the Transcriber's Manual.

It should perhaps be emphasized that the basic structure of the XML schema is rather generic, simply capturing which annotations can be found on a page and where. Although other (meta)data is provided as well, we have consciously limited our editorial interventions, such as linking an annotation to the printed text. If we have done so, this kind of information is provided in several attributes, thus providing users of our tool (and data) with the opportunity to leave out our interpretations of Harvey's reading and annotation practices.

All the XML files have been validated against our schema and other tests which check the internal consistence of the various identifiers of people, books, and places, have been carried out as well. The largest part of our dataset is clean and interesting work on the AoR data therefore can commence!

## Specification of the status of the transcriptions per book:

**Buchanan, *Ane detectioun of the duinges of Marie Quene of Scottes***: All transcriptions have been finalised.

**Buchanan, *De Maria Scotorum regina***: All transcriptions have been finalised.

**Castiglione, *Il libro del cortegiano***: All transcriptions have been finalised.

**Castiglione, *The covrtyer of Covnt Baldessar Castilio***: Initial transcriptions have been done, everything needs to be checked.

**Domenichi, *Facetie, motti, et burle***: Initial transcriptions have been made and most of the checking has been done. However, a number of files have outstanding issues (mostly unclear references to books/people or uncertainties about translations.

**Guicciardini, *Detti et fatti piacevoli et gravi***: Most transcriptions have only been checked partially (in most cases the non-marginal notes, i.e. the marks, symbols, and underscoring still need to be checked). There are also several remaining issues, similar to the Domenichi.

**Freigius, *Paratitla***: All transcriptions have been finished, but there are a couple of remaining issues (in files: 00000005.xml, 00000081.xml, 00000088.xml)

**Frontinus, *The strategems***: Initial transcriptions have been done, everything needs to be checked.

**Livy, *Ab urbe condita***: All the transcriptions have been made, but around 35% of them still need to be checked.

**Machiavelli, *Art of warre***: All transcriptions have been made and checked. Several of them need further revision though.

**Melanchton, *Selectarum declamationum***: All transcriptions have been finished, but there is one remaining issue (file: 00000004.xml).

**Olaus, *Historia de gentibus septentrionalibus***: All transcriptions have been finished, but there are a couple of files with remaining issues.

**Smith, *De recta & emendata linguæ Anglicæ***: All transcriptions have been finished, but there are several with remaining issues (mostly regarding translation and illegibility of annotations).