AOR Phase 1 Transcription Data Summary and Analysis

There are a total of 5,468 individual image files in the AOR corpus (1,146, 21% of that total, comprising images of Livy's *Ab urbe condita* alone). For each image bearing a visible sign of Harvey's manuscript interventions within the printed text—be it a verbal marginal comment, a non-verbal mark or element of Harvey's person symbolic system of annotation, or simple underlining of the printed text—the three transcribers (the AOR postdoctoral researcher and two research assistants) jointly produced a transcription file in XML format. There are 2,501 of these XML files, each associated with its image file counterpart.

Table 1.1 shows how these transcription files are distributed among the 12 volumes in the AOR Phase 1 corpus.

Table 1.1 Number of transcription files per book in corpus

Title book	**Number of files**
Buchanan, *Ane detectioun of the duinges of Marie Quene of Scottes*	50
Buchanan, *De Maria Scotorum regina*	17
Castiglione, *Il libro del cortegiano*	91
Castiglione, *The covrtyer of Covnt Baldessar Castilio*	389
Domenichi, *Facetie, motti, et burle &* Guicciardini, *Detti et fatti piacevoli et gravi*	421
Freigius, *Paratitla*	60
Frontinus, *The strategems*	224
Livy, *Ab urbe condita*	652
Machiavelli, *Art of warre*	238
Melanchton, *Selectarum declamationum*	133
Olaus, *Historia de gentibus septentrionalibus*	121
Smith, *De recta & emendata linguæ Anglicæ*	105
Total	**2,501**

The XML files comprise several "elements", each element corresponding to a type of annotation as described above: marginalia (a verbally written note); underlining (words and passages underlined or highlighted in some way); and a non-verbal mark or symbol. Additional tagging protocols were developed, as well, to deal with more exceptional cases of annotations such as visual drawings; numerals (passages marked up, for example, "1, 2, 3"); and errata (where the annotator appears to have corrected or specifically altered the printed text). In addition, we created the ability to quantify the specific number of words written by Harvey as marginalia, and the number of printed words underlined by Harvey within each book and across the full corpus of books.

As detailed in Table 1.2, below, the 2,501 transcription files contain a total of 102,627 individual elements. More specifically, they contain a total of 80,009 words of manuscript annotations and 226,710 underlined words, comprising a total body of annotated materials amounting to 306,719 words. To our knowledge, this dataset constitutes the largest body of information ever gathered together in a systematic, machine-readable form from manuscript annotations in early printed books.

Table 1.2 General overview of corpus

Title book	Marginalia	Underline	Mark	Symbol	Drawing	Numeral	Errata	*Total*	Marginalia_ words	Underline words	*Total*
Buchanan, *Ane detectioun…*	13	635	221	4	0	0	5	***878***	178	2089	***2267***
Buchanan, *De Maria Scotorum regina*	8	48	3	0	0	0	2	***61***	44	167	***211***
Castiglione, *Il libro del cortegiano*	92	116	48	0	0	0	6	***262***	718	393	***1111***
Castiglione, *The covrtyer of Covnt Baldessar Castilio*	397	11981	5118	633	2	29	20	***18180***	2605	44457	***47062***
Domenichi, *Facetie, motti..* & Guicciardini, *Detti et fatti..*	3094	10876	5820	693	1	13	11	***20508***	39095	33288	***72383***
Freigius, *Paratitla*	50	468	331	31	0	13	2	***895***	1728	1045	***2773***
Frontinus, *The strategems*	869	3721	2549	138	0	18	3	***7298***	10351	11864	***22215***
Livy, *Ab urbe condita*	851	25919	10317	1144	1	66	54	***38352***	21080	95779	***116859***
Machiavelli, *Art of warre*	214	7288	4036	92	0	51	76	***11757***	2630	27393	***30023***
Melanchton, *Selectarum declamationum*	45	1129	327	13	1	8	9	***1532***	429	3206	***3635***
Olaus, *Historia de gentibus septentrionalibus*	53	1552	293	18	0	20	3	***1939***	798	4935	***5733***
Smith, *De recta & emendata linguæ Anglicæ*	42	687	224	6	0	3	3	***965***	353	2094	***2447***
Total	***5728***	***64420***	***29287***	***2772***	***5***	***221***	***194***	***102627***	***80009***	***226710***	***306719***

Table 1.2 therefore provides a fairly objective overview of the reader's interaction with the printed texts in the Phase 1 Harvey corpus. It also reveals several interesting reading patterns undertaken by Harvey across these books. For example, both of Buchanan's books may be regarded as lightly annotated, relative to the rest of the corpus, and yet in Buchanan's *Ane Detectioun*, Harvey underlined far more words of the printed text than were underlined in the second Buchanan book, *De Maria Scotorum*. Although further analysis needs to address the relative length of a book, and the average number of words captured per annotation tag, applied statistical analysis of such major variations within and across books in the corpus (in addition to further exploration of the relationship of common themes that appear both in underlined printed texts and allied marginal annotations) may provide new ways to evaluate

more fundamental elements of historical reading practices and strategies than have been possible in a purely analog environment. These analyses may offer a basis for further and perhaps unexpected lines of enquiry and investigation into historical reading practices.

The humanities content team has, however, also determined that such statistical analysis should not be treated as the endpoint of scholarly investigation of the AOR resource, but rather may be used as a tool to support the identification of specific research questions that might not otherwise occur to the researcher in the analog environment of the library reading room. That research may lead to the identification of demonstrable, even statistically significant relationships only made clear through the analysis of the data extracted from the books in the XML transcription and encoding process, and lead to additional qualitative as well as quantitative scholarly inquiries.

It is also important to note, in the context of the data summarized below in Table 1.2, that the the large majority of printed texts and marginal manuscript annotations in the Phase 1 Harvey corpus were produced in languages other than the English, most notably Latin, and in several modern European vernacular languages (Italian, French, etc.). This repeatedly demonstrates Harvey's general practice of inscribing annotations into books in the same language in which they were originally printed. The XML schema includes a <language> tag to identify which language is being used where. There are 6,243 language tags, of which 1,283 (20.5%) were used to identify passages in English and 4,388 language tags (70.2%) to identify those in Latin. The number of language tags actually exceeds the number of marginalia tags (5,728), since some marginal notes were written in more than one language, and thus one marginalia tag can consist of several language tags. These data may reveal information of particular use to corpus linguistics scholars and to students of the histories of translation and polyglot composition.

This process of transcription was not a straightforward mechanical process, either: in many cases external textual and historical research was required to confirm a given transcription based on contextual evidence contained outside the books in the corpus, or, in some cases, across books within the Harvey corpus, wherever similarities appeared. This was particularly notable in the case of identifying the specific historical personages he cited, identifying specific texts referenced by him in brief short titles, and particular geographic places and locations cited in reference to the printed texts. The Phase 1 corpus transcription files contain 3,950 people tags, 709 book tags, and 221 location tags, which collectively refer (due to repetitions) to 970 individuals, 341 book titles, and 88 locations.

As Table 1.2 also suggests the dataset created in the course of AOR Phase 1 may rightly be regarded as a "big data" set in the sense that the complexity of the data could never be adequately encompassed by an individual researcher working within the constraints of traditional "analog" scholarship. At the same, the average size (in terms of bytes) of the transcriptions is actually is small, as demonstrated below in Table 1.3, relative to data sets generated, for example, in the sciences. This raises an important point in the broader conversation that brings humanistic content in direct conjunction with quantifiable data; namely, that humanists must speak responsibly about the quantitative size of the data that humanities projects generate. The qualitative arguments and analyses that may be generated within humanistic scholarly discourse should not seek to draw greater authority or preponderance simply from the fact that they sit atop a relatively massive pile of quantifiable data (piles of data that would seem, by contrast, remarkably modest in the natural sciences or in the world of finance). It is incumbent upon humanists to consider foremost the qualitative

aspects of the dataset, since they, and not the actual size of the dataset itself, helps the broader scholarly community to address particular research questions. Both forms of data are exceedingly useful, but simply must be treated proportionally.

Table 1.3 Average size of XML file per book

Title book	**Number of files**	**Average length of files (in bytes)**
Buchanan, *Ane detectioun of the duinges of Marie Quene of Scottes*	50	2293
Buchanan, *De Maria Scotorum regina*	17	1047
Castiglione, *Il libro del cortegiano*	91	1181
Castiglione, *The covrtyer of Covnt Baldessar Castilio*	389	5222
Domenichi, *Facetie, motti, et burle &* Guicciardini, *Detti et fatti piacevoli et gravi*	421	8283
Freigius, *Paratitla*	60	2818
Frontinus, *The strategems*	224	5493
Livy, *Ab urbe condita*	652	6972
Machiavelli, *Art of warre*	238	5414
Melanchton, *Selectarum declamationum*	133	1710
Olaus, *Historia de gentibus septentrionalibus*	121	2360
Smith, *De recta & emendata linguæ Anglicæ*	105	1537
Total	**2501**	**5465**

Table 1.3 demonstrates, further, that the AOR XML dataset, no matter its objective size, may generate a range of meaningful insights. For example, it corroborates a pattern that is also visible in Table 1.2; namely, that some books are much more lightly annotated than others, in terms of the relative number of tags generated and words captured. In table 1.3, this is also demonstrated by the average size of the transcriptions within a given book. A quantitative analysis, in this latter context, may prove helpful, but it cannot formulate the basis of an argument that relies on the seeming "size" of the dataset to carry off or prove by default a qualitative point of analysis on that score. It is the quality of the data, rather, and the ways in which they are structured to support user interface with the data, that constitute the major asset of AOR as a digital asset.

During the first year of AOR Phase 1, transcribers have not only generated comprehensive transcriptions and translations of all manuscript annotations in the corpus of 12 books, but have also checked one another's transcriptions independently in order to ensure the accuracy and quality of the data provided in the encoding process. Because of the relatively large size and complexity of this data within a humanities context, that final, secondary review process for checking transcriptions will require further effort within the second year of Phase 1. Currently there are 969 transcriptions (38.7% of the total number of transcriptions) which still need to be checked in this way, as demonstrated in Table 1.4, below.

Table 1.4 Number of files that remain to be checked in AOR Phase 1, year 2

Title book	**Number of files which need to be checked**
Castiglione, *The covrtyer of Covnt Baldessar Castilio*	389
Domenichi, *Facetie, motti, et burle &* Guicciardini, *Detti et fatti piacevoli et gravi*	119
Frontinus, *The strategems*	224
Livy, *Ab urbe condita*	237
Total	**969**

The items in Table 1.4 reflect those books in the Harvey corpus for which the transcriptions have been completed, but still require independent cross-checking. There are an additional c. 75 remaining questions regarding humanistic content within completed transcriptions of annotations that are still unresolved due to ambiguous or complex translations, the presence of as yet indecipherable, or otherwise unknown non-verbal marks or symbols, as well as particular references to books or historical figures that are still difficult to identify at present. Our goal in AOR Year 2 is to resolve all these remaining tasks of the humanities team, and complete this portion of the annotations work by March 31, 2015.