

Digital Aeschylus

Breadth and Depth Issues in Digital Libraries

Federico Boschetti

CIMeC, University of Trento, Italy
federico.boschetti@unitn.it

Abstract. Digital Libraries can grow along two different dimensions: breadth and depth. In the first case, works of many authors extend the existing collections. In the second case, different editions of the same works and related studies populate a monothematic region of the library. The *Digital Aeschylus Project* is aimed to collect, link and process digital objects based on primary and secondary sources related to the ancient Greek tragic poet.

1 Digital Libraries and Philological Needs

The most complete Greek and Latin corpora of texts, such as the Thesaurus Linguae Graecae (TLG) and the Packard Humanities Institute (PHI) Latin collection, are based on authoritative, most recent critical editions of each classical author. In these collections, only the text established by the editor is digitized, whereas the critical apparatus is omitted. Such approach to the ancient text, just about acceptable for literary and linguistic purposes, is unfeasible for philological studies. In fact, the philologist needs to identify manuscript variants and scholars' conjectures, in order to evaluate which is the most probable textual reading, accepting or rejecting the hypotheses of the previous editors. Furthermore, he or she needs to examine the commentaries, articles and monographs concerning specific parts of the text. Thus, the extension in breadth of the aforementioned collections needs to be integrated by the extension in depth, according to the paradigms of a new generation of digital libraries (see [7] and [18]).

In order to go in depth, philological studies are necessarily focused on single authors, genres or periods, even if they need to find links and parallels in the entire Greek and Latin literature. For this reason, teams of specialists need to share a common infrastructure, as pointed out by [8]).

For instance, the [15] Project is building a cyberinfrastructure to interrelate different philological and archeological projects. The [13] Project, on the other hand, has created a large platform to manage textual variants of Latin texts. The [12] Project, even if it is focused on a single ancient author, has developed a suite of services that can be easily extended to other authors.

The *Digital Aeschylus Project*, <<http://www.himeros.eu/digitalaeschylus>>, aims to provide philologists with a search engine on primary and secondary sources for the study of the Athenian tragic poet's tradition, taking into account

the standards for annotation and textual references that are emerging in the domain of the digital philology.

2 Structure of the Digital Aeschylus Project

The project is structured in modules, concerning the bibliographical catalogation, the acquisition of digital images of the documents, manual transcription and annotation of the most relevant manuscripts and early printed editions, OCR of recent editions, information extraction from the digitized documents. Due to the loose interdependence of the modules, they can be easily developed asynchronously.

2.1 Bibliographical Catalogue

The bibliographical catalogue related to manuscripts, printed editions and studies on Aeschylus aims to extend and integrate the printed repertory edited by [20]. In particular, it will supply the references to the digital resources on Aeschylus provided by large and general purpose digital libraries, such as [9] and [10], or provided directly by the research team, directed by V. Citti, that is working to the new edition of Aeschylus' tragedies.

The catalogue can be considered the roadmap for the digitization process, because it registers which resources are online and which ones are not yet accessible.

2.2 Digital Diplomatic Editions

Digital images of Aeschylean manuscripts and early editions have been collected under the direction of V. Citti and M. Taufer (see [19], for the current status of the acquisition).

The second step concerns the transcription and annotation of the most relevant materials, according to the Text Encoding Initiative guidelines related to manuscripts and early editions, [6]. These digital documents can be automatically collated, in order to identify textual variants and collect them in dynamic critical apparatus. (Techniques for automated collation by multiple alignment algorithms are illustrated in [17]). Furthermore, the different layouts of the manuscripts can be compared, going along with the recent interest for the colometric assessment of the tragic choral parts. In fact, it is demonstrated that the study of the colometry, i. e. the disposition of the strophes in different lines, sheds light on the generation of transmission errors.

Unfortunately, as demonstrated in [16], the optical character recognition on early editions is still unsatisfactory. For this reason, editions printed before the XIX century must be manually transcribed like the manuscripts. Naturally, the transcription is not performed by scratch, but by modification of a digital copy of a recent edition.

2.3 OCR on Recent Printed Editions

OCR can be applied to XIX and XX century critical editions, reaching up to 99% of accuracy on the text and more than 90% of accuracy on the critical apparatus. Anyway, it is important to point out that the critical apparatus is approximately only 5% of the page in editions with minimal information, and approximately 14% of the page for more informative editions.

These performances are obtained by the alignment and merging of three different OCR outputs and the application of an automated system of spell-checking, supported by the evidence of the OCR outputs. After suitable training, both [1] FineReader 9.0 and [14] 0.3 are able to recognize polytonic Greek characters mixed to Latin characters, whereas [2] 4.1 is able to recognize only polytonic Greek. Each OCR engine is more or less reliable for specific characters, and the reliability is evaluated by training sets. The merging system computes the most probable character in each position: the result significantly overwhelms the performances of the single engines.

The details of the system are illustrated in [5] and a similar approach is exposed in [11].

2.4 Information Extraction from the Repertories of Conjectures

Repertories of conjectures register not only the corrections to the ancient text suggested by the editors in their own editions, but also the proposals for emendation contained in commentaries and articles. As illustrated in [3], the repertories of conjectures have a trivial structure: in fact, more than the 90% of the items are constituted by the reference to the verse affected, the text of the conjecture and the name of the scholar that has made the proposal. A parser identifies these chunks of information and an alignment algorithm is applied to find the exact position in the verse where the conjectures is intended to be collocated.

3 Putting all Together

The search engine that will be developed, extending the model of [13], should allow the research of variants and conjectures in their contexts, showing the actual page of the manuscript where the variant is attested or the image of the printed edition where the conjecture was formulated.

4 Conclusions

Digital Aeschylus is an ongoing project focused on the textual tradition of a single ancient author. The first stage concerns the acquisition, digitization and linkage of the materials.

The second stage will concern the application of corpus analysis to the acquired documents. First experiments on the current corpus are promising, as illustrated in [4].

References

1. Abbyy FineReader Homepage, <http://www.abbyy.com>
2. Anagnostis Homepage, <http://www.ideatech-online.com>
3. F. Boschetti: Methods to Extend Greek and Latin Corpora with Variants and Conjectures: Mapping Critical Apparatuses onto Reference Text. Proceedings of the Corpus Linguistic Conference (27-30 July 2007)
http://ucrel.lancs.ac.uk/publications/CL2007/paper/150_Paper.pdf
4. F. Boschetti: Gli Spazi Semantici del Greco Antico. (to appear in Quaderni Urbinati 2008)
5. F. Boschetti, M. Romanello, A. Babeu, D. Bamman, G. Crane: Improving OCR Accuracy for Classical Critical Editions. (to appear in ECDL 2009)
6. L. Burnard, S. Bauman: TEI P5 – Guidelines for Electronic Text Encoding and Interchange. Oxford (2008)
<http://www.tei-c.org/Guidelines/P5>
7. G. Crane, D. Bamman, L. Cerrato, A. Jones, D. Mimno, A. Packel, D. Sculley, G. Weaver: Beyond Digital Incunabula: Modeling the Next Generation of Digital Libraries. 10th European Conference on Research and Advanced Technology for Digital Libraries, volume 4172 of Lecture Notes in Computer Science, 353-366, Springer (2006)
8. G. Crane, B. Seales, M. Terras: Cyberinfrastructure for Classical Philology. *Digital Humanities Quarterly*, 3, 1, 1–27 (2009)
9. Internet Archive Homepage, <http://www.archive.org>
10. JStor Homepage, <http://www.jstor.org>
11. W.B. Lund, E.K. Ringger: Improving Optical Character Recognition through Efficient Multiple System Alignment. (to appear in JCDL 2009)
12. Multitext Homer Homepage,
http://chs.harvard.edu/chs/homer_multitext
13. Musisque Deoque Homepage, <http://www.mqdq.it>
14. OCRopus Homepage, <http://code.google.com/p/ocropus>
15. Perseus Project Homepage,
<http://www.perseus.tufts.edu/hopper/opensource>
16. S. Reddy, G. Crane: A Document Recognition System for Early Modern Latin. *Chicago Colloquium on Digital Humanities and Computer Science: What Do You Do With A Million Books*, Chicago, IL (2006).
17. M. Spencer, C. Howe: Collating texts using progressive multiple alignment. *Computer and the Humanities*, 37, 1, 97–109 (2003)
18. G. Stewart, G. Crane, A. Babeu: A New Generation of Textual Corpora. *JCDL 2007*, 356–365 (2007)
19. M. Tauffer: Stato del New Repertory of Conjectures on Aeschylus e della Collezione di Manoscritti Eschilei. (to appear in Quaderni Urbinati 2009)
20. A. Wartelle: *Bibliographie Historique et Critique d'Eschyle et de la Tragédie Grecque*, 1518-1974. Paris (1978)