Massimo Manca, Linda Spinazzè, Paolo Mastandrea, Luigi Tessarolo, Federico Boschetti

Musique Deoque: Text Retrieval on Critical Editions

This paper aims at illustrating the main features of the *Musisque Deoque Project*, which provides a fully freely searchable archive of Latin poetry equipped with critical apparatus. The first part explains how variants are mapped on the reference edition and the second part illustrates the web interface to retrieve sequences of words taking into account possible variants.

1 Introduction

The *Musisque Deoque Project* (MQDQ) aims at creating a digital archive of Latin poetry, from its origins to the late Italian Renaissance, equipped with critical apparatus and various exegetical and linguistic information. This project is focused on the study of synchronical and diachronical intertextuality as illustrated, e.g., in Cicu (2005). For this reason, we give strong attention to formal and material aspects of the text that actually played a relevant role in the poetical tradition. The fixed text of printed critical editions, aimed at the reconstruction as close as possible to the lost originals, provides just a snapshot of the tradition, which is intrisically dynamic, and gives to the modern reader a distorted image of what an ancient text was in fact.

Fully searchable digital collections currently available are based on traditional critical editions, which are, as we just said, authoritarian texts; this authoritarianism is emphasized by the conversion from printed text to database, because usually the critical apparatus is cut away and there is no way for the reader to check a variant different from the one the editor put in the main text, often *dubitanter*, simply because he *had* to choose a variant. Limiting lexical searches to editor's choices drives unavoidably both to false positives and false negatives, which need to be verified back on printed critical editions. False positives are due to possibly wrong emendations made by modern and contemporary scholars, provided by the text retrieval systems among the genuine occurrences, whereas false negatives are the likely variants excluded by editors biased by prejudices against specific linguistic and stylistic phenomena (such as the short-term repetiton, systematically emended by philologists of the last centuries).

The purpose of Musisque Deoque is to overcome these limitations, retrieving not only the word keys quoted in the reference edition, but also the variants lying in the critical apparatus. In this way, further knowledge on the accomplished itinerary – from ancient operas during the subsequent ages until the Humanism and the Renaissance – can emerge.

2 Background

Musisque Deoque is the result of the continuous evolution of projects focused on the digitization of the Latin poetry: *Almae Latinitatis Bibliotheca* (Classical Latin Poetry), *Poetria Nova* (Medieval Latin Poetry) and *Poeti d'Italia in Lingua Latina* (Humanist and Renaissance Italian Poetry in Latin). Along the decades, additional information has been encoded to the text-only documents related to metrical genres, to biographical data of the authors and other information and, consequently, features have been added to the search engines available on CD or online (in particular, *Poeti d'Italia*: http://www.poetiditalia.it and *Musisque Deoque*: http://www.mqdq.it).

Important projects that deal with digital variants have been developed in the last decades: see, in particular, Calabretto and Bozzi (1998) and Calabretto et al. (2005). These projects are focused on the collation of manuscripts and are aimed to provide tools that help the philologist to check variants on the images of the manuscripts or to produce an automated collation of digital diplomatic editions. On digital philology and Medieval texts, see Stella and Ciula (2007).

The Perseus Project stressed out the importance of a cyberinfrastructure for the classical studies able to deal also with variants. See, for instance, Crane (2009) and Crane (2010).

Peter Robinson, the promoter of *Interedition* and *Virtual Manuscript Room*, considers the process of editing digital editions as a collaborative enterprise: see Robinson (2010) and Babeu (2011); see also Price (2009) and McGann et al. about digital scholarship. In this perspective, the main purpose of *Interedition* is offering a sort of public, social and sharing context in order to improve, compare and discuss first of all the tools for digital scholarly edition publishing. Very similar ideas about the future development of digital scholarly editions are asserted by Gabler (2010).

MQDQ does not aim to the *constitutio textus* nor offers new protocols for publishing digital editions; its goal is rather to offer a tool to study the literary influences among the tradition. The ideal end-user of MQDQ is a scholar interested in analyzing the *Fortleben* and the mutual relationship of texts at a more deeper level than the one allowed by the common authoritative databases.

Even if MQDQ takes into account the theoretical models and the practices to represent variants, as expressed in recent contributions, such as Boschetti (2007), McGann (2010), Gabler (2010), Marotti (2010), May (2010), Mandell (2010), the main goal of MQDQ project is to achieve a very extensive database, which includes almost all the Latin poetry with a wide range of variants.

MQDQ is a work in progress and its features and improvements have been illustrated in several conferences and articles, such as Zurli and Mastandrea (2009), Manca (2009) and Mastandrea (2011).

3 Encoding Text and Critical Apparatus

Musisque Deoque is based on dynamic repertories of texts and critical apparatus. On one hand, the text of a classical work is a faithful transcription of the text established by the editor of the most authoritative printed critical edition currently available and only in few cases it is digitized from the text established by scholars on articles, Ph.D. theses, etc. Pages are scanned, OCR is performed, and skilled operators select only text boxes excluding the critical apparatus. Manual corrections made by the operators are reviewed by the project managers and their collaborators.

On the other hand, many critical editions of the same work are examined by skilled operators. They prepare the digital *conspectus codicum* and they insert the variants that they consider the most relevant for the study of the textual tradition.

The concept of "significant variant" isn't so subjective as it sounds. Manca (2009) defines "significant variant" a "*lectio* we can credit to the author himself, or to an editor, but more often introduced by readers or copists still in the ancient phase of the tradition, and which may bring to new perspectives in intertextual researches". A variant that trace the path of the textual tradition must be considered significant, even when it is an error from a metrical, syntactic, pragmatic or encyclopedic point of view. For example, a variant such as *Gallia omnis est divisa in partes quattuor* would be surely rejected in any traditional critical apparatus; but if this mistake would have been elaborated by the literary tradition, one should accept it in a *corpus-oriented* apparatus. For a more realistic case, see the success, in the scriptural tradition, of the 'wrong' expression *Nabuzardan princeps coquorum* over the 'correct' *Nabuzardan princeps militiae* in Manca (1999). In the MQDQ enviroment it is significant also a variant *deterior* or *facilior*, completely useless for the *constitutio textus*, if this different reading somehow spreaded itself in literature. Usually corrupted variants, cases of *scriptio continua* or wrong division are not significant into our archive.

In order to enrich very quickly the existent database of latin texts with more variants for every single works, the first technical effort was to build a user-friendly tool that permits to scholars unskilled in IT but very competent philologists to become MQDQ-operators. The MQDQ operator works with a cross-platform sofware written in Java called MQDQ2. The philologist that creates new digital editions with MQDQ may decide to download and modify the text present in the pre-existent database or to replace it with another plain text (i.e. without tags). The user have to initialize the text for adding apparatus information through a sequence of dialogue boxes (Fig. 1): in this phase the operator writes the header of xml file for the apparatus and decides how to create the *conspectus codicum*.

The table of manuscripts and bibliography can be encapsulated in the same XML file of the apparatus, or can be saved as an indipendent XML file: it is up to the operator in the first phase to choose the preferred method. The wizard that guides the operator that starts working with a new text offers the author a choice to build a new *conspectus codicum* or to share an existent one, usually taken from a different section of the same

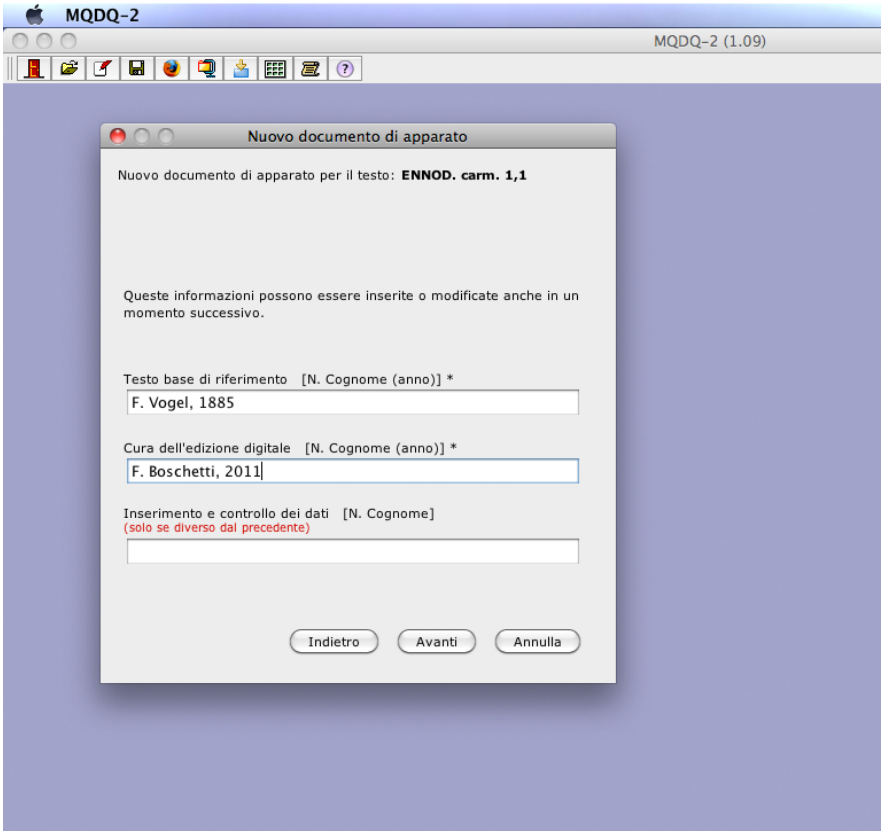

Figure 1: Initializing the interface 1/1

author. Very often, indeed, it is useful to choose the same table of *codices* among several sections of the same work, or different works of same author(s) (Fig. 2).

After the preliminary operations, the user accesses the main page of the application, where he can build a *conspectus codicum et uirorum doctissimorum* (Fig. 3) and add the variants or other kind of paratextual notes (Fig. 4). The system is enough flexible to allow the operator to cope with the almost endless problems of information representation in a text.

So, variants and conjectures registered in different critical editions are inserted in the new digital critical apparatus, and each textual variant is mapped on the reference edition.

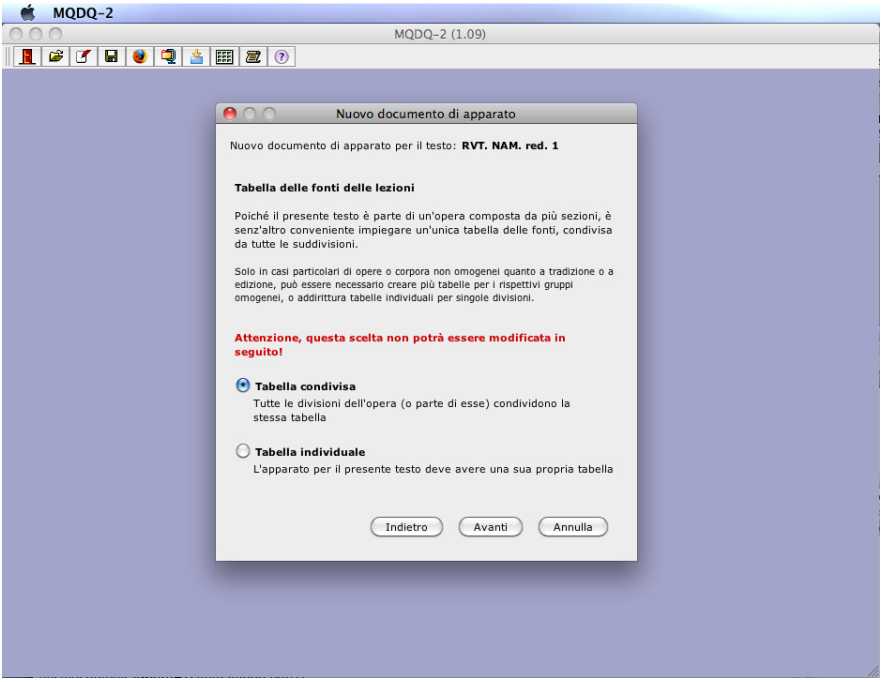

Figure 2: Initializing the interface 1/2

The structure of the back-end is transparent to the operators, which are not supposed to be skilled in XML annotation, but the hidden structure is worthy of mention.

In order to decouple text and apparatus, the text is fully segmented at the word level. Each word or punctuation mark receives a unique identifier, which is used to map the variant on the correct position in the context of the verse, as illustrated below:

```
<line id="l38" name="36" type="verse">
        <word id="w239">patrem</word>
        <word id="w240">probaui</word>
<punctuation id="w241" space="post">,</punctuation>
        <word id="w242">gloriae</word>
<word id="w243">feci</word>
<word id="w244">locum</word>
<punctuation id="w245" space="post">.</punctuation>
</line>
```

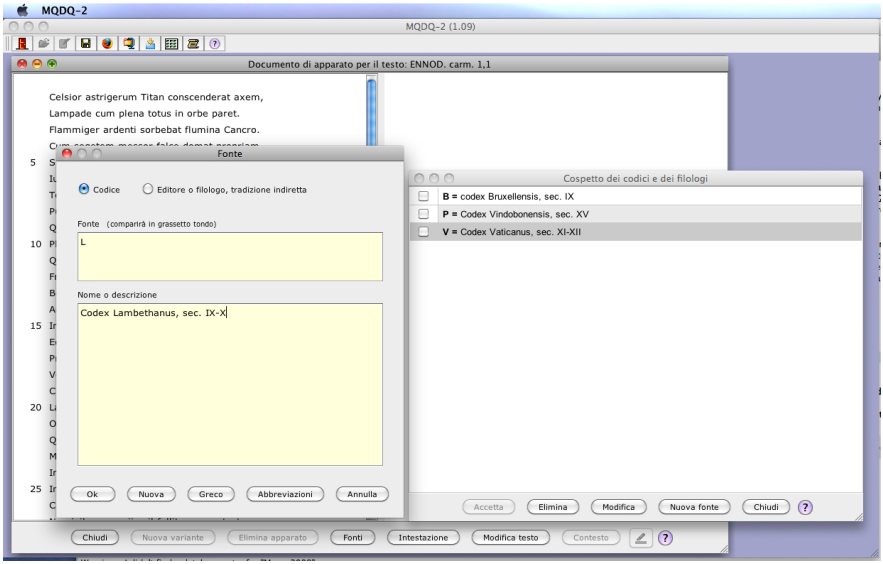

Figure 3: Creating conspectus codicum

```
<line id="139" name="37" type="verse">
<word id="w246">qua</word>
<word id="w247">Sol</word>
<word id="w248">reducens</word>
<word id="w249">quaque</word>
<word id="w250">deponens</word>
<word id="w251">diem</word>
</line>
```

The *conspectus codicum et virorum doctissimorum* is built in a separate file (or encapsulated in the same app.xml file, see above), encoding the names of the editions from which the information is extracted, *sigla*, description of manuscripts, and the name possibly with reached bibliographical information about scholars that proposed conjectures. As seen above, this *conspectus* collects information from many critical editions of the same work. The editor of the printed edition and the editor of the electronic edition are mentioned within the `head` tag. Each source has an identifier, which is unique for the metadata related to a specific work (e.g. *s1* for the code **R** that contains Seneca's tragedies). But this identifier can be equivalent to the identifiers related to other works (e.g. when a miscellaneous manuscript is mentioned in different ways by the scholars).

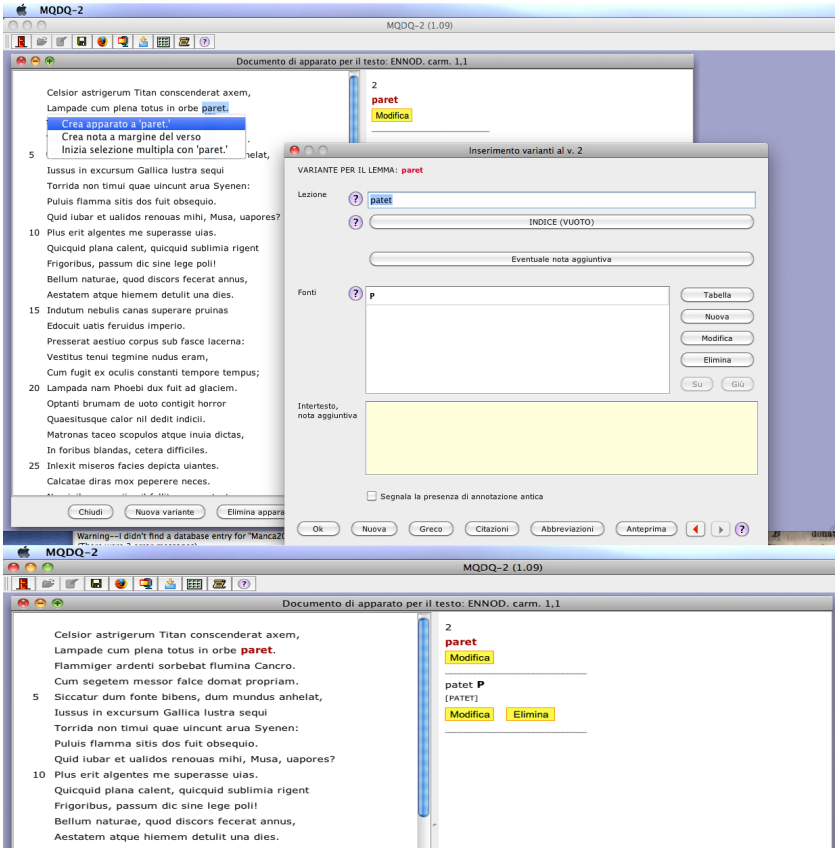

Figure 4: Creating variant readings

A conversion table allows the suitable correspondence. Each source has also a type, which can be `cod` (for a manuscript) or `auth` (for a scholar that suggested a conjecture), as shown below:

```
[...]
    <head>
        <editor>O. Zwierlein (1986)</editor>
        <e_editor>G. C. Musa (2008)</e_editor>
    </head>
    <links>
        <text>xml-app/SEN-hefu-001-txt.xml</text>
    </links>
```

```
    <conspectus>
        <source id="s1" type="cod">
            <name>R</name>
            <explication>Ambrosianus G 82 sup., saec. V in., 5 foll. rescripta</explication>
        </source>
        <source id="s2" type="cod">
            <name>E</name>
            <explication>Laurentianus Plut. 37. 13
                (&amp;quot;Etruscus&amp;quot;), saec. XI ex., foll. 165</explication>
        </source>
        [...]
        <source id="s67" type="auth">
            <name>Commelinus</name>
            <explication>H. Commelinus apud Scriuerium</explication>
        </source>
        [...]
```

After the preamble that contains the *conspectus*, chunks affected by multiple readings are registered. Each chunk is uniquely identified and it contains a reference to the textual positions where insertions, deletions, substitutions or translations are required by alternative readings.

The reading accepted by the editor of the reference edition is marked as `pos`, whereas the other variants are marked as `var`. The text of the reading is inserted and words are indexed with the positions in the verse. In case of addition of text, positions can be identified by decimal numbers. In case of deletion of text, the `operation` tag is used. `note` is used to insert unstructered information from the printed critical apparatus, such as evaluations of scholars (e.g. *dubitanter...*); the encoding of a chunk of information is shown below:

```
[...]
<chunk id="i12" nameVerse="37" idRef="w246w247w248w249w250">
        <reading type="pos">
                <reading>qua Sol reducens quaque deponens</reading>
                <source idSources="s2">
                    <operation></operation>
                    <note></note>
                </source>
        </reading>
        <reading type="var">
                <reading>aperitque thetis qua ferens titan</reading>
                <index idRef="w246">APERITQVE</index>
                <index idRef="w247">THETIS</index>
                <index idRef="w248">QVA</index>
                <index idRef="w249">FERENS</index>
                <index idRef="w250">TITAN</index>
                <source idSources="s16">
                    <operation></operation>
                    <note></note>
                </source>
        </reading>
</chunk>
[...]
```

As said above, the resulting XML document is not visible to the operators, mostly graduate or Ph.D. students of Latin literature, which insert data via a front-end interface

that allows the selection of text on the reference edition and the completion of related forms with the information about variants. This back-end system produces a coherent XML code among all the operators and, as no XML training is necessary, a scholar with no previous knowledge of tagging may be operative with only a short two-hours briefing.

4 Querying

The simplest function of the web interface is the retrieval of word sequences. MQDQ inherits the metrical metadata encoded in the previous projects directed by P. Mastandrea and L. Tessarolo, such as Poeti d'Italia in Lingua Latina. It is possible to find words in special positions of the verse (in particular the beginning and the end), to filter specific metres (e.g. dactylic metres) and also to search words inside the extra-text, the sender of a letter, the speaker of dialogues etc. (see Fig. 5).

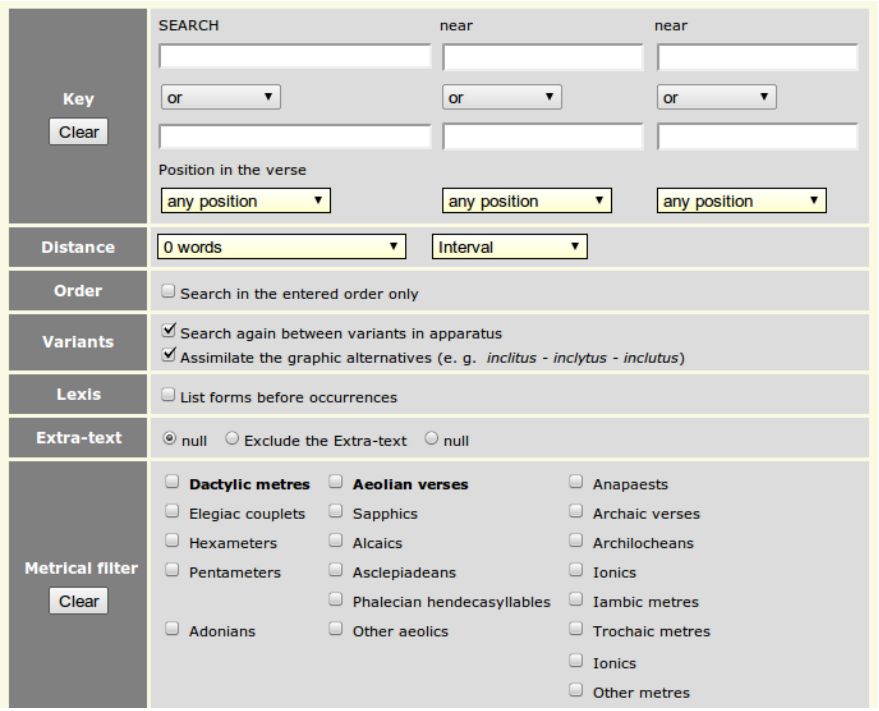

Figure 5: Querying Interface

The query mask allows to search word sequences (and graphic alternatives, such as words written with `-dc-` or with `-cc-`) not only in the reference editions, but also in the collection of variants, collocated in the correct context of the verse. That means that a sequence constituted by a word of the reference edition followed by a word registered in the critical apparatus can be recognized as adjacent and retrieved.

Figure 6: Variants

Searching for *amore leget*, it is possibile to retrieve also the variants *amore legat* or *amor releget*. Searching for *Captus amor*, it is possible to find the occurrence even if one word was originally recorded in the reference edition and the other was originally encoded only in the critical apparatus (see Fig. 6).

5 Conclusion

In conclusion, Musisque Deoque provides tools both to build digital critical editions and to query a large database of variants, which are mapped on reference editions. MQDQ is focused on the study of the intertextuality, and for this reason is based neither on digital diplomatic editions of manuscripts nor on the mere text of traditional critical editions, but on a selection of relevant variants recorded in printed critical editions. With increasing digital material such as manuscripts and ancient edition on the Web, the team of MQDQ are now ready to work in the direction of interoperability, expanding the traditional intertextuality among ancient texts to the new intertextuality that the power of the Internet nowadays offers, according to the developments illustrated in Spinazzè (2011).

References

Babeu, A. (2011). *Rome Wasn't Digitized in a Day: Building a Cyberinfrastructure for Digital Classicists*. CLIR Reports.

Boschetti, F. (2007). Methods to extend greek and latin corpora with variants and conjectures: mapping critical apparatuses onto reference text. In *Proceedings of the Corpus Linguistics Conference*.

Calabretto, S. and Bozzi, A. (1998). The philological workstation bambi (better access to manuscripts and browsing of images). *J. Digit. Inf.*, 1(3).

Calabretto, S., Bozzi, A., Corradini, M. S., and Tellez, B. (2005). The eumme project: towards a new philological workstation. In *ELPUB*.

Cicu, L. (2005). *Le api il miele la poesia. Dialettica intertestuale e sistema letterario greco-latino*. Roma.

Crane, G. (2009). Cyberinfracstructure for classical philology. *DHQ*, 3(1).

Crane, G. (2010). Give us editors! re-inventing the edition and re-thinking the humanities. In *Online Humanities scolarship: the Shape of things to Come*, <http://cnx.org/content/m34316/latest>. Mellon Foundation, Jerome McGann, ed.

Gabler, H. W. (2010). Theorizing the digital scholarly edition. *Literature Compass*, 7(2):43–56.

Manca, M. (1999). Nabuzardan princeps coquorum. una lezione vulgata oltre la vulgata. *Quaderni del Dipartimento di Filologia, Linguistica e Tradizione Classica (Università degli Studi di Torino*, (13):491–498.

Manca, M. (2009). Database and corpora of ancient texts towards the "second dimension": theory and practice of musisque deoque project. In P. Anreiter, M. K., editor, *Computational Linguistics and Latin Philology. 15th Colloquium on Latin Linguistics.*, pages 697–702.

Mandell, L. (2010). Special issue: 'scholarly editing in the twenty-first century'– a conclusion. *Literature Compass*, 7(2):120–133.

Marotti, A. F. (2010). Editing manuscripts in print and digital forms. *Literature Compass*, 7(2):89–94.

Mastandrea, P., editor (2011). *Nuovi archivi e mezzi d'analisi per i testi poetici. I lavori del progetto Musisque Deoque, Venezia 21-23 giugno 2010*. Hakkert.

May, S. W. (2010). All of the above: The importance of multiple editions of renaissance manuscripts. *Literature Compass*, 7(2):95–101.

McGann, J. (2010). Electronic archives and critical editing. *Literature Compass*, 7(2):37–42.

McGann, J., Gabler, H. W., Wolfson, S. J., Pratt, L., Curran, S., Marotti, A. F., May, S. W., Ezell, M. J. M., O'Donnell, D. P., and Mandell, L. *Literature Compass*, (2):134–144.

Price, K. (2009). Edition, project, database, archive, thematic research collection: What's in a name? *Digital Humanities Quaterly*, 3(3).

Robinson, P. (2010). Editing without walls. *Literature Compass*, 7(2):57–61.

Spinazzè, L. (2011). Risalire alle fonti: dall'edizione mqdq ai testimoni manoscritti. In Mastandrea, P., editor, *Nuovi archivi e mezzi d'analisi per i testi poetici*, pages 59–71. Hakkert.

Stella, F. and Ciula, A., editors (2007). *Digital philology and medieval texts. Atti del seminario, Arezzo 2006*. Pacini.

Zurli, L. and Mastandrea, P., editors (2009). *Poesia latina, nuova E-filologia. Opportunità per l'editore e per l'interprete. Atti del convegno Internazionale. Perugia 13-15 settembre 2007*. Hedder Editrice e Libreria.