Collaborative Multimedia Platform for Computational Philology

CoPhi Architecture

Angelo Mario Del Grosso
Istituto di Linguistica Computazionale "A. Zampolli"
Consiglio Nazionale delle Ricerche
Pisa, Italy
angelo.delgrosso@ilc.cnr.it

Federico Boschetti
federico.boschetti@ilc.cnr.it

Abstract— **This paper aims at illustrating a collaborative and modular web platform in the domain of digital and computational philology. The proposed work deals with parallel multilingual and multimedia resources. Two case studies are discussed in order to show the flexibility of the designed platform. The reusability of the components in different projects is achieved by abstract modeling and through the application of effective design patterns. The platform deals with textual resources and associated multimedia content, which can be retrieved by the metadata and shown in parallel (e.g., the page image of a manuscripts and the related transcription). The library of components will distribute under GPL 3.0 license and available at https://github.com/CoPhi.**

Keywords-computational philology; digital philology; multilingualism; parallel multimedia; enterprise systems.

I. INTRODUCTION

A general introduction provides an overview on initiatives and projects related to the specific domain of digital philology. The proposed methodology, illustrated in the second section, is based on the application of design patterns. Two pilot projects constitute the results shown in the third section: the former is an application for digital epigraphy and the latter is an application to manage manuscripts. Finally, conclusions are discussed, in order to point out the flexibility and reusability of the system.

Current trends in digital and computational philology are focused on the implementation of collaborative environments, in order to provide the international scholarly community with a suitable infrastructure to share and reuse scientific products, such as digital critical editions, commentaries, linguistic and stylistic analyses, annotations on manuscript page images, descriptions of archaeological and epigraphical artifacts, etc.[13-17][19][23]

The main efforts to achieve this aim are (1) the standardization of data formats for literary and philological studies[2][16][19][23], (2) the modelling of domain ontologies[4-6], (3) the interconnection with disciplines focused on text-bearing objects, such as Digital Epigraphy[13][23] or artifacts mentioned in literary works, such as Digital Archaeology and, eventually, (4) the management of data by web-services through common protocols, such as OAI-PMH [1].

The standardization of data formats for literary and philological studies is the mission of the Text Encoding Initiative (TEI [2]). The TEI provides XML schemes and guidelines for text, extra-text, and para-text encoding with bibliographical, linguistic and philological meta-information.

Domain ontologies are provided by various institutions. Europeana [3], a platform for information and knowledge exchange in the domain of Digital Humanities, has formalized a data model that is becoming a de-facto standard (EDM: Europeana Data Model [4]).

The CIDOC-CRM [5], on the other hand, provides the conceptual reference model in the domain of Digital Archaeology and a joint effort between CIDOC-CRM and the Functional Requirements for Bibliographic Records has created FRBRoo [6]: an ontology intended to link bibliographical and museographic information.

Web-services allow the data exchange among working groups distributed world-wide. In the field of classical literature and philology, Bamboo [7] and Interedition [8] aim at providing web-services to make critical editions in collaborative environments.

The Perseus Project [9] (Tufts University) is the leading initiative that provides scholars with the suitable cyberinfrastructure: Philologist, powered by Son of Suda (SoSol [10]), and Alpheios [11] are the web applications that allow the version-controlled editing of texts and linguistic analyses associated to them. The identification of textual units is formalized by the Canonical Text Service (CTS [12]), which associate a URN (Uniform Resource Name) to every word of any specific edition.

As shown above, standard data formats on one hand and web-services on the other hand, highly promote the interoperability. But in the field computational philology it is necessary to improve also the software reusability. Whereas libraries and API for information retrieval, such as Lucene or linguistic analysis, such as LingPipe exist and are

 ISBN: 978-1-61208-265-3

maintained, libraries devoted to the specific field of computational philology are wanted: computational linguistics analyzes a single text flow associated to single linguistic analyses (e.g., syntactic and semantic analyses), whereas computational philology must deal with multiple versions of the same text (due to variants in the manuscripts or conjectural emendations provided by the scholars) and multiple interpretations at each level of analysis (due to the disagreement of authoritative scholars recorded in several commentaries along the centuries).

The main purpose of our work is the constitution of a library of components (the CoPhi Beans library) focused on philological activities, such as the alignment of complex textual objects (e.g., the alignment of variants according to their semantic similarity, not only according to the edit distance of the inflected forms), the extension of levels of analysis (e.g., metrical and colometrical analysis) or editing and retrieval of multiple, concurrent annotations. Furthermore, the linkage of textual resources to multimedia sources, such as the manuscript page images, must be taken into account in the new paradigm of philology in the digital age, which pays increasing attention to the disintermediation between the philologist and the (digital representation of) the primary sources.

Due to the abstract modeling and the modular design, our library and the platform based on, even if it is still in alpha version, is already used in a number of national and international projects devoted to manage parallel multimedia resources, such as text, image and music combined together.

Three main areas are involved for developing a platform based on the CoPhi Beans library able to deal with such spread needs:

- Acquisition of resources by Optical Character Recognition (OCR) or information extraction and document transformation from semi-structured resources to structured ones (ETL: Extract, Transform and Load);
- Text processing and indexing;
- Collaborative Enterprise Application designing and developing.

Eventually, we are designing and developing components, modules and plug-ins for a collaborative enterprise web-based system in order to build a suitable environment to analyze, on one hand, manuscripts and printed documents and, on the other hand, to produce new critical editions.

II. METHODOLOGY

Flexibility and reuse is achieved by a modular and abstract design of the components.

The core of the platform is: (1) the view resources component, (2) the search and index component, (3) the analysis component, (4) the comment and collaborative component (5) the editorial component.

The architecture of our platform is based on the MVC (Model View Controller) pattern, which separates the business model from the GUI (Graphical User Interface) and from the objects devoted to the behavioral aspects. Fig. 1 shows the class model design involved in commenting results and the design model involved in the analysis process.

Comments written by the users in natural language and micro-annotations automatically produced by morpho-syntactic parsers must be editable versioned and searchable. Furthermore Textual components can be composite with linguistic analyses and multimedia resources (images, audio, etc.) in a flexible way, at different levels of granularity (single words, sentences, paragraphs, documents, etc.). In order to achieve this goal, compound patterns have been used: (1) Composite (2) Typed Relationship (3) Factory Method.

The whole system has been developed according to the Java Server Faces 2 specification (JSR-314) and according to the general JSR-316 specification (Java Enterprise Edition 6). Persistence is obtained using a native XML-DB, eXist, which stores and retrieves documents encoded in TEI or other XML compliant documents.

The system, currently in alpha release, is producing promising results and several deployments in real projects, such as the Res Gestae Divi Augusti Web App and the Saussure Web App, which will be illustrated below.

III. RESULTS

A. *Res Gestae Divi Augusti Web Application*

The Res Gestae Divi Augusti Web Application handles the Mommsen's edition of the well known bilingual epigraph (Fig. 2). The aligned texts are shown in parallel and the granularity of the information is flexible: the units can be paragraphs for a coarse alignment, words for the morphological analysis and characters for the annotation of the status on the stone, such as readable or unreadable.

The comment component allows scholars to record exegetical annotations, commenting fragments on a selected chunk of text, which can be labeled. Index and Search component is suitable for advanced text retrieval based on metadata produced both by automatic processes and by scholar studies. For example in this project it is possible to perform queries for chunks in both languages (Greek and Latin) and word status on the available support (e.g., attested or conjectured).

B. *A digital edition of F. de Saussure's Manuscripts*

A prototype of the digital edition of Saussure's texts, based on a representative selection of his manuscript images and transcriptions, has been the focus of a Research Programs of Relevant National Interest (PRIN2008). This project has been a test for evaluating the platform with text and image resources. The text has been extracted from semi-structured electronic documents and transformed in a TEI-compliant format. Fig. 3 shows the component that manages parallel resources, in order to browse and search both texts and images. The links are referred to annotations that author of the manuscripts has done and to which editor refers in the critical apparatus. As shown in Fig. 1, the platform is based

 ISBN: 978-1-61208-265-3

on component aggregation (e.g., comments), or on the extension of components (e.g., indexes). Many indexes (one per language used in Saussure's quotations) are required in this project. Consequently, the component is able to handle indexes and concordances for each language.

The comment component is useful to enhance collaborative annotations and arrange canonical linkage between selected text, such as named entity and authoritative resources on the web infrastructure.

IV. CONCLUSION

The new paradigm of computational philology in the current web generation is the collaborative philology, based on crowdsourcing both for software development and for the acquisition and generation of data. Even if many libraries of components have been implemented for computational linguistics, which can be reused in the domain of philology, computational philologists must afford specific issues due to multiple readings of the same text and multiple interpretations of the same reading. Furthermore, collaborative philology is open to other disciplines, such as palaeography, codicology, musicology, which heavily involve multimedia. For this reason we have illustrated how we are developing a library of java components that constitutes the core of a web platform, focused on the domain of collaborative philology to deal with texts and related multimedia sources. As seen above, the modularity of the system promotes the reuse of philological components in many applications. This aim is achieved both by aggregation and by extension.

Future work gives priority to the multimedia aspects of the philological studies. The same framework used to create a new textual edition from many corrupted witnesses can be applied to restore a sound track from many low quality, noisy or incomplete recordings of the same concert. APIs developed by third parties devoted to specific tasks, such as sound track alignment, will be studied and interfaced to our platform.

In conclusion, collaborative philology is experiencing a double transition: on one hand it is moving from single, desktop applications towards web applications, web services and distributed programming, on the other hand it is widening its attention from text to multimedia.

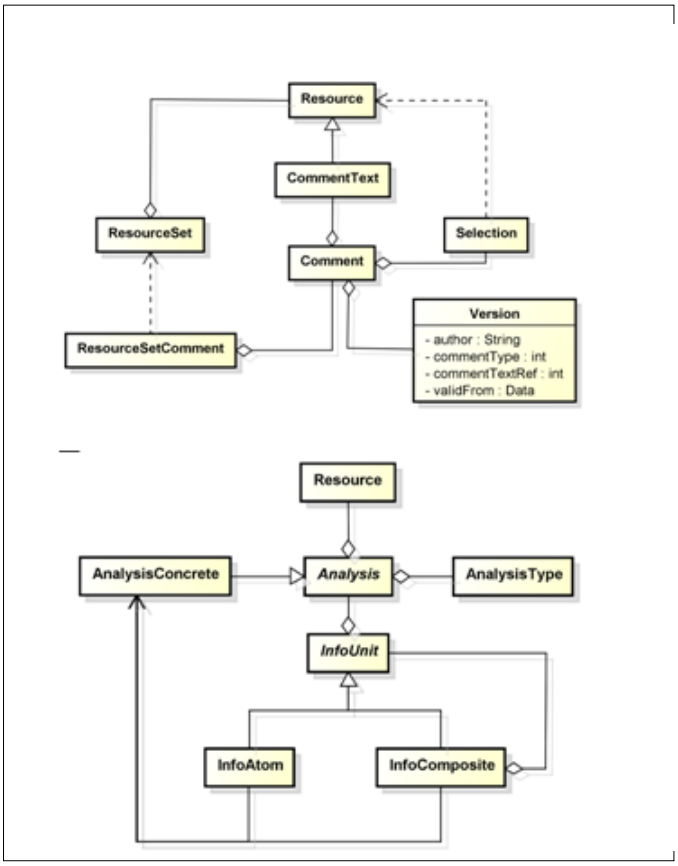

Figure 1. Class Model Diagram for Comment and Analysis components

 ISBN: 978-1-61208-265-3

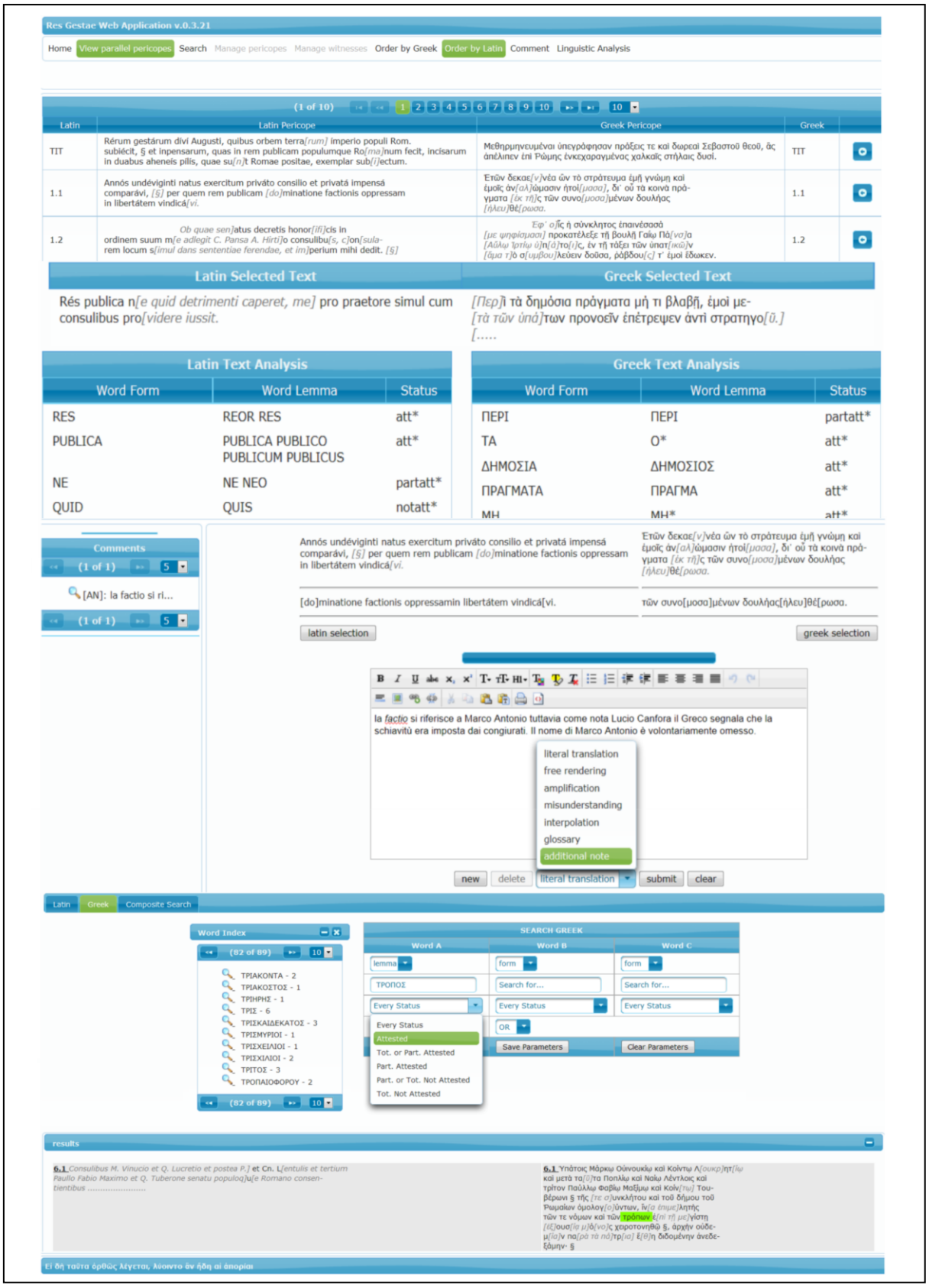

Figure 2. Res Gestae Divi Augusti deployment

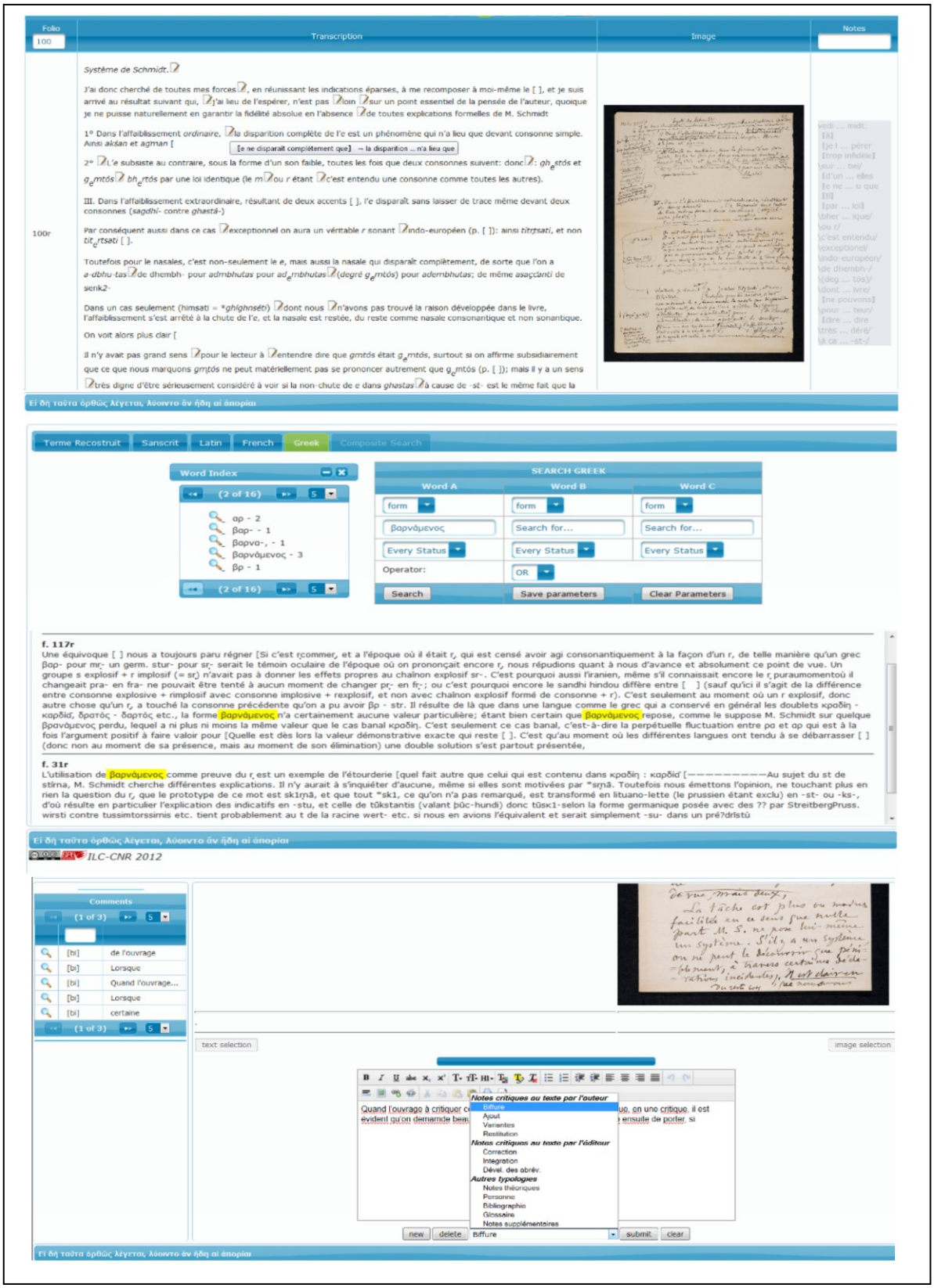

Figure 3. A digital edition of F. de Saussure's Manuscript

"REFERENCES"

[1] http://www.openarchives.org/pmh [retrieved February, 2013]

[2] http://www.tei-c.org/index.xml [retrieved February, 2013]

[3] http://www.europeana.eu/portal/ [retrieved February, 2013]

[4] http://pro.europeana.eu/edm-documentation [retrieved February, 2013]

[5] http://www.cidoc-crm.org/ [retrieved February, 2013]

[6] http://www.ifla.org/node/928 - http://www.cidoc-crm.org/frbr_inro.html [retrieved February, 2013]

[7] http://www.projectbamboo.org/ [retrieved February, 2013]

[8] http://www.interedition.eu/ [retrieved February, 2013]

[9] http://www.perseus.tufts.edu/hopper/ [retrieved February, 2013]

[10] http://idp.atlantides.org/trac/idp/wiki/SoSOL/Overview [retrieved February, 2013]

[11] http://alpheios.net/ [retrieved February, 2013]

[12] http://www.homermultitext.org/hmt-doc/cite/index.html [retrieved February, 2013]

[13] A. Berra, "Exploitation de la matière épigraphique dans un espace numérique, Edition savante et humanités numériques". http://philologia.hypotheses.org/648 [retrieved February, 2013].

[14] A. Bozzi, M. M. Morales, and M. Rufino, "Imago et umbra: Programma di digitalizzazione per l'Archivio storico della Pontificia Universita Gregoriana: criteri, metodi e strumenti," in Digitalia, Anno V, Numero 2, Roma, 2010, pp 79-99.

[15] A. Bozzi, V. Sandrucci, "Uno strumento al servizio dell'archiviazione, lo studio, l'edizione e l'interrogazione di documenti digitali," in Carmen Alén Garabato, Mercedes Brea, Xosé Afonso Álvarez (a cura di), Quelle linguistique romane au XXIe siècle?, Paris: L'Harmattan, Langue et parole, 2010, pp. 27-40.

[16] A. Bozzi, "Edizione elettronica e filologia computazionale", in A. Stussi (a cura di), "Fondamenti di critica testuale", Il Mulino Manuali, Bologna, 2006, pp. 207-232.

[17] A. Bozzi, "Towards a philological workstation," in Revue informatique et statistique dans les Sciences humaines, XXIX, 1993, pp. 33-49.

[18] S. Burbeck, "Applications Programming in Smalltalk-80TM: How to use Model-View-Controller (MVC)," 1992 http://st-www.cs.illinois.edu/users/smarch/st-docs/mvc.html. [retrieved February, 2013.

[19] G. Crane, B.Seales, and M. Terras, "Cyberinfrastructure for Classical Philology," Digital Humanities Quarterly, 3 (1), 2009 URL: http://www.digitalhumanities.org/dhq/vol/3/1/000023/000023.html [retrieved February, 2013.

[20] M. Fowler, "Analysis Patterns: Reusable Object Models". Menlo Park, Calif. ; Harlow : Addison Wesley. 1996.

[21] E. Gamma, R. Helm, R.Johnson, and J. Vlissides, "Design Patterns: Elements of Reusable Object-Oriented Software". Reading, Mass: Addison-Wesley, 1995.

[22] G.Hohpe, B. Woolf, "Enterprise Integration Patterns: Designing, Building, and Deploying Messaging Solutions". Boston: Addison-Wesley, 2004.

[23] M. Lamé, V. Valchera, and F. Boschetti, "Epigrafia digitale. Paradigmi di rappresentazione per il trattamento digitale delle epigrafi," Epigraphica, LXXIV vol. 1-2, 2012, pp. 331-338

[24] M. McCandless, E. Hatcher, and O. Gospodnetić, "Lucene in action", Manning, 2010.

[25] G. Stewart, G. Crane, and A. Babeu: A New Generation of Textual Corpora. JCDL 2007, pp. 356–365.

 ISBN: 978-1-61208-265-3