CLARIN-IT: texts, documents and new contexts

Federico Boschetti
ILC "A. Zampolli" CNR, Pisa
& VeDPh, Venezia, Italy
federico.boschetti@ilc.cnr.it

Angelo Maria Del Grosso
ILC "A. Zampolli" CNR
Pisa, Italy
angelo.delgrosso@ilc.cnr.it

Riccardo Del Gratta
ILC "A. Zampolli" CNR
Pisa, Italy
riccardo.delgratta@ilc.cnr.it

Francesca Frontinini
ILC "A. Zampolli" CNR
Pisa, Italy
francesca.frontini@ilc.cnr.it

Monica Monachini
ILC "A. Zampolli" CNR
Pisa, Italy
monica.monachini@ilc.cnr.it

Abstract

In recent years, CLARIN has increasingly broadened its interest from linguistic resources to textual resources relevant to digital humanists. This new and attractive scenario requires new technologies for texts, variants, and digital representations of primary sources, their contexts, and complex relationships. VeDPH in Venice, CNR-ILC-CoPhiLab, and ILC4CLARIN in Pisa collaborate on DH projects. Together, they are working on extracting text from manuscript page images, annotating historical graffiti on georeferenced images, and identifying text in digital images of paintings and sculptures.

1 Introduction

The acronym CLARIN is an acronym for Common Language Resources and Technology Infrastructure, and it reflects the fact that the principal community to which it addressed its activities was originally composed of linguists and computational linguists. However, Language resources, such as dictionaries or wordnets, e.g. "Ancient Greek WordNet" (Bizzoni et al., 2014)[1], and textual resources, such as literary or documentary corpora, e.g. "Cretan Institutional Inscriptions" (Vagionakis et al., 2022)[2] are complementary instruments to study the immaterial cultural assets of a civilization. The new definition of CLARIN as "the research infrastructure for language as social and cultural data[3]" is consistent with this new and more extended vision of language and its contexts.

In the last years, CLARIN made an effort to meet the requirements of the Digital Humanities and Museums-Libraries-Archives communities (Del Fante et al., 2022). For this reason, CLARIN has broadened its boundaries toward the digital representation of cultural artifacts, in particular towards the digital representation of text-bearing objects, such as papyri, inscriptions, and manuscripts.

Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR) are the necessary links between the digital representation of primary sources (i.e. facsimile images) and the conveyed textual content. CLARIN-IT is exploring the most suitable open OCR and HTR tools and services to be integrated into its infrastructure. At this stage, *eScriptorium* (Kiessling et al., 2019)[4] seems to be the best trade-off between openness of the licenses and recognition accuracy. *eScriptorium* exploits the IIIF protocol to seamless import facsimiles provided by authoritative digital archives (such as e-Codices[5] or Ambrosiana[6]) and digital libraries (such as Gallica[7] or the Bodleian Digital Library[8]). CLARIN-IT can help the Italian community of digital humanists in three ways. First, CLARIN-IT can host the manifests created by scholars and provide a permanent identifier (PID) through a handler service. Second, it can

[1] To deepen into the Ancient Greek WordNet see http://hdl.handle.net/20.500.11752/ILC-56

[2] To deepen into the Cretan Institutional Inscriptions project see http://hdl.handle.net/20.500.11752/OPEN-548

[3] The new definition is claimed on CLARIN HomePage: https://www.clarin.eu/

[4] The eScriptorium git repository can be found at https://gitlab.com/scripta/escriptorium

[5] e-Codices digital archive can be found at the following web address https://www.e-codices.unifr.ch/it

[6] Ambrosiana digital archive can be found at the following web address https://www.ambrosiana.it

[7] Gallica digital library can be found at the following URL https://gallica.bnf.fr

[8] Bodleian Digital Library can be found at the following URL https://digital.bodleian.ox.ac.uk

create a special interest group to work on the integration of OCR or HTR data expressed in standard formats, such as ALTO[9], and metadata expressed through the IIIF manifests[10]. Third, the K-Centre devoted to Digital and Public Textual Scholarship (DiPtext-KC[11]), is planning workshops and seminars about the IIIF best practices.

2 Collaboration between VeDPH, CNR-ILC-CoPhiLab and ILC4CLARIN

In the Italian scenario, Computational Linguistics and Digital Humanities have a long story of entanglements and separations (Buzzetti, 2019), of methodological sharings and high specialization in knowledge subdomains (Montemagni, 2013). Among the others, we focus our attention on a specific case of collaboration. The Venice Centre for Digital and Public Humanities[12] (VeDPH) of the Department of Humanities at the Ca' Foscari University of Venice has been working in synergy with the Collaborative and Cooperative Philology Lab[13] (CoPhiLab) of the CNR-Institute of Computational Linguistics "A. Zampolli"[14] (CNR-ILC) and with the B-Centre ILC4CLARIN[15] since its founding in 2019 (Fischer et al., 2023).

3 HTRoman and HTRogène (Italian Section)

VeDPH takes part to the projects HTRoman and HTRogène, lead by the University PSL (Paris Sciences et Lettres) and funded by Biblissima+[16]. The aim of the projects is the enlargement of the HTR-United[17] (Chagué et al., 2021) collection of accurate transcriptions of samples from Medieval manuscripts with heterogeneous layouts, written in different scripts for various Romance languages: ancient French, Occitan, Catalan, Castillan, Tuscan, and Venetian. VeDPH, supported by CNR-ILC, is working on the manuscripts produced in Italy. For HTRoman, a team of two proof-readers and a supervisor accessed eScriptorium[18]. The web platform integrates the following functionalities: a) image acquisition through uploading or through the IIIF protocol; b) layout analysis; c) text recogntion (through Kraken[19]); d) proof-reading; e) creation of a new model or of a fine-tuned model. Like the other national sections of the project, also the data related to the Italian section are available online under an open access license[20].

4 VeLa

Venezia Libro Aperto (VeLa, Venice Open Book) is a DH project lead by the Department of Humanities of Ca' Foscari University devoted to the digitization of the historical graffiti of Venice (De Rubeis, 2008), with the high priority of Ducal Palace. The project, currently funded by Biblissima+, involves VeDPH, MUVE[21], SABAP-VE-MET[22], CESCM[23] and CNR-ILC.

The project consists in the creation of a shared georeferenced database of all the graffiti of the Doge's Palace in Venice, from the 15th to the 20th century. Multiple transparent layers (according to different original hands and chronology) with the hand-drawn transcriptions are superimposed over the high resolution images of the graffiti.

Contextual metadata (related to place, shape, material, etc.) and textual data (related to transcriptions, named entities, etc.) will be encoded through VeLaDSL, a domain-specific language easily convertible in XML-TEI/EpiDoc to ensure the interoperability with other Biblissima+ projects.

[9] The ALTO XML document format is described by the following specifications https://www.loc.gov/standards/alto/
[10] The last API specifications are described at the following URL https://iiif.io/api/presentation/3.0/
[11] To deepen into the k-centre website see https://diptext-kc.clarin-it.it
[12] https://www.unive.it/pag/39287
[13] https://cophilab.ilc.cnr.it/
[14] http://www.ilc.cnr.it/
[15] https://ilc4clarin.ilc.cnr.it/
[16] https://biblissima.fr/
[17] https://htr-united.github.io/
[18] https://escriptorium.inria.fr/
[19] https://github.com/mittagessen/kraken
[20] https://github.com/HTRomance-Project/medieval-italian
[21] https://www.visitmuve.it/en/home/
[22] https://www.soprintendenzapdve.beniculturali.it/
[23] https://cescm.labo.univ-poitiers.fr/

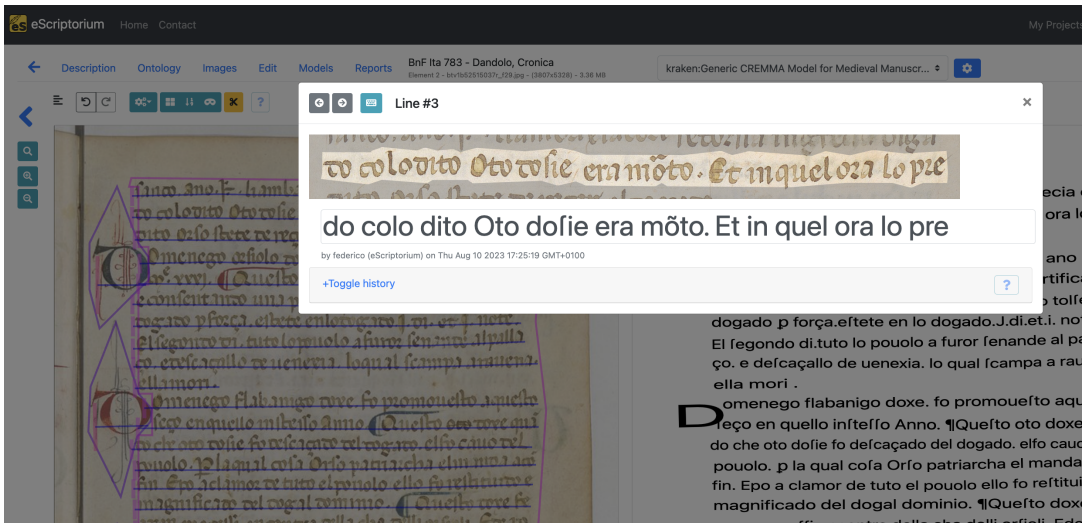

Figure 1: eScriptorium

5 Galleria Borghese

The project for the virtual museum of the Galleria Borghese in Rome (De Vincentis & Critelli, 2023) allows the navigation of highest resolution images, with a 360 degrees perspective, of the rooms of the gallery. Paintings and sculptures can be zoomed by exploiting the framework developed by the IIIF community. Images can be annotated according to the Web Annotation Data Model[24]. An interesting aspect of the project is the possibility to annotate regions of the digital images containing written texts (for example inscriptions on the basements of the sculptures or cartouches and scrolls within the paintings to transcribe the texts and to make them searchable and linkable to other textual sources. Part of data, navigation and annotation tools will be hosted in the new data center of the H2IOSC[25] consortium, located in Pisa.

6 On the possible integration in CLARIN-IT

The integration illustrated in Vagionakis et al., 2022 represents a model for other comparable DH projects. For the projects outlined in this contribution, we will follow the described strategy: a) the use of the repository of CLARIN-IT to describe both the data and the tool; b) the provisioning of the services through the CLARIN-IT servers, and c) a GitHub repository with data and software. This strategy requires to address some points and raises different questions.

In this section, we focus on (i) licenses for both data and tools; (ii) hardware and software requirements, and (iii) versioning.

Licenses (i) are of fundamental importance for a). As CLARIN, we can describe images and texts from the projects if they have been properly licensed. At this stage, licenses are not defined yet, but we may assume the images (at least a substantial subset of the entire collection) will be licensed under a CC-BY-SA[-NC]. This allows us to describe the (subset of) data in the ILC4CLARIN repository without defining a specifying license and specific access policies to the resources. The model described in Vagionakis et al., 2022 applies a total decoupling between data accessed by the CLARIN repository and data accessed by the application. We may follow the same strategy and limit the provisioning of the offered services to the hosting of such services, b). In such a case, ILC4CLARIN acts as the host of the IIIF servers but does

[24] https://www.w3.org/TR/annotation-model/
[25] https://www.h2iosc.cnr.it/

not interfere with the licences and access policies. The decoupling is important to dimension hardware and software as well, (ii). The ILC4CLARIN center will be dramatically improved during H2IOSC, both in terms of storage and GPUs. However, the ILC4CLARIN will be only a component of the H2IOSC project. b) allows us to host the IIIF services on different servers. Finally, (iii) and c) define a methodology and a workflow: we require developers to use GitHub as a repository for images and software. Every time a new release is available, a new version of the images in the ILC4CLARIN repository is submitted as well as a new version of the provided services. In this way, we guarantee the replicability: researchers can access previous version of data and software and replicate their experiments.

7 Conclusion

The collaboration of VeDPH in Venice with CNR-ILC-CoPhiLab and ILC4CLARIN in Pisa is an opportunity to work for the integration of language and textual technologies with image technologies in order to have a wider perspective on language as cultural data. Furthermore, the collaboration allows us to better address the following difficulties: a) to ensure the long term preservation and maintenance to projects based on the linkage of textual and visual resources and b) to constantly share the know-how among CLARIN, CNR and university, even when the projects receive small funding and consequently the turnover of human resources devoted to them is frequent.

References

Bizzoni, Y., Boschetti, F., Del Gratta, R., Diakoff, H., Monachini, M., & Crane, G. (2014). The making of Ancient Greek WordNet. *Proceedings of the 9th Annual Conference of LREC*.

Buzzetti, D. (2019). The Origins of Humanities Computing and the Digital Humanities Turn. *Humanist Studies & the Digital Age*, *6*(1), 32–58. https://doi.org/10.5399/uo/hsda.6.1.3

Chagué, A., Clérice, T., & Romary, L. (2021). Htr-united: Mutualisons la vérité de terrain! *DHNord2021-Publier, partager, réutiliser les données de la recherche: les data papers et leurs enjeux*.

De Rubeis, E., Flavia; Banterla. (2008). Scrivere sui pavimenti, scrivere sui muri: Materiali originali e riuso architettonico. In *Monasteri in europa occidentale (secoli viii - xi). topografia e strutture* (pp. 477–487). http://opac.regesta-imperii.de/id/1335049

De Vincentis, S., & Critelli, M. (2023). Mappare il museo in IIIF. Una combinazione di deep zoom e VR360 per la Galleria Borghese di Roma. *Proceedings of AIUCD2023*.

Del Fante, D., Frontini, F., Monachini, M., & Quochi, V. (2022). CLARIN-IT: An Overview on the Italian Clarin Consortium After Six Years of Activity.

Fischer, F., Boschetti, F., Del Grosso, A. M., Montefusco, A., Mancinelli, T., & Macchiarelli, A. (2023). Sinergie fra VeDPH e CNR-ILC in termini di condivisione della conoscenza e sostenibilità dei progetti digitali. *DH.22*. https://doi.org/10.48255/9788891328342.08

Kiessling, B., Tissot, R., Stokes, P., & Ezra, D. S. B. (2019). Escriptorium: An open source platform for historical document analysis. *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, *2*, 19–19.

Montemagni, S. (2013). DH@ILC. In M. Agosti & F. Tomasi (Eds.), *Collaborative Research Practices and Shared Infrastructures for Humanities Computing. Proceedings of revised papers of the 2nd Annual Conference of the Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD), Padova* (pp. 101–114). CLEUP.

Vagionakis, I., Del Gratta, R., Boschetti, F., Baroni, P., Del Grosso, A. M., Mancinelli, T., & Monachini, M. (2022). 'Cretan Institutional Inscriptions' Meets CLARIN-IT [ISBN: 978-91-7929-444-1 ISBN: 1650-3686]. In F. de Jong & M. Monachini (Eds.), *Selected Papers from the CLARIN Annual Conference 2021* (p. 189). CLARIN ERIC. https://doi.org/https://doi.org/10.3384/9789179294441