"Tea for two": the Archive of the Italian Latinity of the Middle Ages meets the CLARIN infrastructure

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

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

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

Marina Buzzoni
ALIM, Università Ca' Foscari
Venezia, Italy
mbuzzoni@unive.it

Paolo Monella
ALIM, Università degli
Studi di Palermo, Italy
paolo.monella@unipa.it

Roberto Rosselli Del Turco
ALIM, Università degli
Studi di Torino, Italy
roberto.rossellidelturco@unito.it

Abstract

This paper presents the Archive of the Italian Latinity of the Middle Ages (ALIM) and focuses, particularly, on its structure and metadata for its integration into the ILC4CLARIN repository. Access to this archive of Latin texts produced in Italy during the Middle Ages is of great importance in providing CLARIN-IT and the CLARIN community, at large, with critically reliable texts for the use of philologists, historians of literature, historians of institutions, culture and science of the Middle Ages.

1 Introduction

The Archive of the Italian Latinity of the Middle Ages – in Italian, Archivio della Latinità Italiana del Medioevo (ALIM) – is an Italian national research project aimed to provide free online access to a large number of Latin texts produced in Italy during the Middle Ages. ALIM makes an unprecedented contribution to the studies not only of Latin, but also of culture and science at the basis of our Wester European society. The general aim of the paper is to allocate ALIM within the framework of CLARIN-IT and CLARIN at large. Section 2, shows how ALIM may contribute to fill an important gap in textual sources: query searches run on the Virtual Language Observatory for Latin-related resources demonstrate that no resource with the features and potentialities of ALIM is currently available. The technical description of the internal structure and metadata of the Archive is discussed in Section 3 while the strategy for the integration of the ALIM archive into the ILC4CLARIN repository is discussed in Section 4. Finally, ALIM's benefits for the CLARIN-IT research directions and for the CLARIN community are presented.

2 Classics resources in CLARIN

The Italian CLARIN (CLARIN-IT) consortium[1] has strong interest in the field of Digital Classics, which still suffers from shortage or restricted availability of lexical resources for historical languages.

Within the CLARIN-IT consortium, the collaboration between the Centre for Comparative Studies "I Deug-Su", Department of Philology and Literary Criticism at the University of Siena – DFCLAM[2] and the data centers mostly concerns the study of methods and services to offer online secure access to some digital archives of literary and historical texts: among them, ALIM (the Archive of the Italian Latinity of the Middle Ages), hosted by the University of Siena, http://alim.unisi.it/il-progetto/), is the largest digital library of the Italian Latinity including both literary and documentary texts.

Evidence shows that the CLARIN data centers do not offer resources such as ALIM. The faceted-search functionality of the the Virtual Language Observatory (VLO), performed combining *Latin text resource* and *Middle Ages*, returns 53 records only (124 are returned by the query *Latin* combined with the adjective *medieval*). These data are essentially images of manuscripts; no XML-TEI texts seem to be available by using these search keys. A further query with XML as free text, Latin as language, and text or corpora as resource type returns about 1300 records, mostly consisting in Treebanks or in documents coming from EUROPEANA[3].

Having ALIM in CLARIN is, thus, very important for both CLARIN and the community of users. ALIM will offer high-quality resources because of the following reasons: (i) the resources are digitized by domain experts; (ii) the large number of resources that cover a wide historical period; (iii) a strong organization dedicated to the maintenance; (iv) all the offered resources are TEI encoded.

3 ALIM: history, goals, structure

ALIM is an archive of medieval Latin texts composed in the Italian area between the $8th$ and $15th$ century. It originated as a UAN (Unione Accademica Nazionale) project in the Nineties and was later supported by the national Ministry of Education. Its original aim was twofold: to make medieval Latin literature texts openly available and to provide a textual corpus serving as a basis to create a new dictionary of medieval Latin in its Italian variety. The latter goal explains a unique feature of ALIM: it does not only include literary sources, but also collections of documentary texts. ALIM is therefore divided into two sections: "Fonti letterarie" and "Fonti documentarie". While the majority of texts are drawn from printed editions, some are new, born-digital editions[4].

3.1 From ALIM1 to ALIM2, from ALIM2 to CLARIN: text and metadata

Until 2016, ALIM was hosted by the servers of the University of Verona, Italy[5] and its texts had procedural markup, based on simple HTML markers. We shall refer to this version as "ALIM1".

In 2016, the current version of the archive ("ALIM2") was launched. The migration process involved the following tasks: (1) building a new open source software TEI XML-based digital library infrastructure and publishing it in the servers of the University of Siena[6]; (2) re-encoding text markup and metadata in TEI XML P5.

Task 1 was realised in collaboration with the external IT company Net7[7] and completed in 2016/17, when the ALIM2 website was launched. Task 2 involved a longer process, still ongoing, curated by the "équipe di codifica" (E. Ferrarini, P. Monella and R. Rosselli Del Turco) with the aim of gradually improving the level of formalization and the granularity of text markup and metadata.

[1] The composition of the Italian Consortium cf. (Nicolas et al., 2018) is available at `http://clarin-it.it/en/content/consortium`

[2] Prof. Francesco Vincenzo Stella

[3] `http://www.europeana.eu`

[4] More information on the history and scientific objectives of ALIM, with further bibliography, are in (Alessio, 2003); (Buzzoni and Rosselli Del Turco, 2016, par. 7.1.2); (Ferrarini, 2017) and (D'Angelo and Monella, 2019).

[5] `http://www.alim.dfll.univr.it/`

[6] `http://alim.unisi.it/`

[7] `https://www.netseven.it/`

In the current version of the archive, each literary text is encoded as a TEI XML P5 file with a `<TEI>` root element, while in the documentary section each TEI XML file includes a whole volume of a documentary collection[8], has a `<teiCorpus>` root element and includes each individual document in a separate `<TEI>` element. In the latter case, both `<teiCorpus>` and `<TEI>` have their own `<teiHeader>` with metadata respectively regarding the whole collection and the individual document.

In ALIM2 TEI XML files for literary texts deriving from the initial export from ALIM1 (labelled as "encoding level ALIM2_0"), much metadata was included in unstructured `<note>` elements of the TEI. Also, most texts lacked any TEI structural markup such as `<div>`. In 2017/18, literary texts were gradually upgraded to "encoding level ALIM2_1" thanks to the work of ALIM collaborator Chiara Casali on metadata integrity, and of Jan Ctibor on metadata encoding and structural markup. Jan Ctibor's activity was brought forth in the framework of a collaboration agreement between ALIM and the *Corpus Corporum*[9], the largest full-text repository for Latin (163 M words). The current policy of ALIM requires that all new texts included to the archive must be encoded at "level ALIM2_2": this includes markup of work titles, quotes, speeches, person or place names.

The archive also includes born-digital scholarly editions directly based on handwritten medieval witnesses, whose encoding level is labelled as ALIM2_3[10].

The ALIM project provided CLARIN-IT with the TEI headers of the XML files in the archive, at the highest available encoding level, to extract metadata from them.

4 ALIM in CLARIN-IT

4.1 Structure for ALIM data into ILC4CLARIN repository

As described in Section 3, the ALIM digital library is arranged into two complementary sections: *Fonti Letterarie (Literary Sources)* and *Fonti Documentarie (Documentary Sources)*. The former is a collection of single documents (about 350), while the latter is a collection of 50 corpora that groups about 6455 texts. Since ALIM keeps these two resources separated, we decided to mirror this structure in the ILC4CLARIN repository. We created two collections, *Literary Sources* and *Documentary Sources*, under the *OPEN* community[11]. This structure is important for, at least, two reasons. The first one is that researchers accustomed to ALIM find in the ILC4CLARIN repository the same structure they are used to; the second one is connected with the VLO. In section 2, we briefly mentioned the faceted-search of the VLO. Well, one of such facets is exactly the collection (in the repository) the data come from. The ALIM data are retrieved from the VLO using either "fq=collection:ALIM+Literary+Sources&fqType=collection"or "fq=collection:ALIM+Documentary+Sources&fqType=collection"[12].

4.2 Population of the repository with ALIM data

The about 350 *Literary Sources* have complete descriptive metadata, although period, author and title are often debated in the scholarly community and therefore tentative in the collection. Author names have two issues: the actual authorship attribution and alternative Latin spellings of the name. Titles too are not always standardised, and the very identification of the "work", as well as of the composition period, is problematic. However, each of these metadata fields has a value in ALIM (for the author, it can also be "Anonimo"). The 50 corpora of *Documentary Sources* group 6455 small documents. For these small documents the metadata set differ from *Literary Sources*', since they do not represent a creative work by an author. For example, private documents are actually written by a notary, but their "author" is the stakeholder (the person who buys, sells etc.), while charters are created by a public institution.

[8] E.g.: *Codex diplomaticus Cavensis*, volume 1: `http://alim.unisi.it/dl/fonte_documentaria/7381`.

[9] `http://www.mlat.uzh.ch/MLS/`

[10] See `http://alim.unisi.it/collection/nuove-edizioni-editiones-principes-e-prime-trascrizioni/` for a list of such editions. In general, on markup levels see the *Manuale di codifica dei testi ALIM in TEI XML* in `http://alim.unisi.it/documentazione/`

[11] Since ILC4CLARIN uses the clarin-dspace repository, we have used the terminology community and collections. For clarity, collections are nested into communities.

[12] At the time of writing, only the *Literary Sources* have been imported into the ILC4CLARIN production repository. The items are available at `https:/dspace-clarin-it.ilc.cnr.it/repository/mlui/handle/000-c0-111/130`.

As a consequence, we decided to completely import *Literary Sources* metadata into the repository, but, at the same time, to describe only the 50 corpora of *Documentary Sources*, without importing the whole amount of data (even if technically possible).

The ratio behind this decision is related to the ALIM organization again. As noted in Section 3.1, the TEI version of each document in literary sources has its own `<teiHeader>`, corresponding to the `TEI` root element, that can be parsed. While for documentary sources, the most informative `<teiHeader>` is extracted from `<teiCorpus>`, for literary sources metadata are extracted from the header of each files' `<TEI>` element.

Given the large number of items to describe in the repository, we decided to use the import functionality of the repository[13] to batch-load the items. Since this procedure is unsupervised, as far as the content of the items is concerned, we decided to manually create a prototypical item, export it, and automatically clone it. In this way, every item is syntactically correct and can be safely imported into the repository.
In details: (i) we took one document from literary sources and one from documentary works and kept them as prototypes; (ii) we carefully created a submission, mapping the elements of the `<teiHeader>` into the fields of the submission form of the repository; and (iii) once the internal workflow of metadata quality is passed, we exported the item.
The exported item is an archive which contains the following metadata files: **metadata_local.xml**, **dublin_core.xml**, and **metadata_metashare.xml**. All of them are populated with data extracted from elements of the `<teiHeader>`. The ALIM research team checked sample metadata from the CLARIN archive and verified that they correspond to those included in the TEI headers of the ALIM XML files and to the general project information pertaining to the archive. Before concluding, let's add that the official URL of the ALIM project (in our case, `http://it.alim.unisi.it/`) is contained in the **dublin_core.xml** files, while **metadata_local.xml** files contain the *demo URL* This mapping enforces our decision to describe the `<teiCorpus>` instead of describing every single document in the corpus. Literary Sources have a clear URL where the document resides: for example, the "Dialogus" by Gerius Aretinus is available at `http://it.alim.unisi.it/dl/resource/194`. By contrast, Documentary Sources point to URLs that report the whole corpus. For example, the "Codex diplomaticus Cavensis - 01" is available at `http://it.alim.unisi.it/dl/fonte_documentaria/7381`. On the web page, a JavaScript allows the user to jump to the desired documents, such as the 27^{th} document, whose internal URL is `http://it.alim.unisi.it/dl/fonte_documentaria/7381#doc_27`. Unfortunately, '#' is a reserved character[14] which separates information sent to server from client side actions, and no data transmitted as part of the URL must contain it.
The complete mapping guide, the scripts, and XSLT style sheets are available at `https://github.com/cnr-ilc/alim2clarin-dspace`.

4.3 Versioning

The ILC4CLARIN repository implements the versioning of the described items. Indeed, it is always possible to add to the repository a new item as "new version of" a previous one. The versioning of the items on the repository should be consistent with the one on the ALIM digital library. The latter allows contributors to replace the XML-TEI of a literary work or documentary collection with a new one, including changes in the text or in the metadata. The ALIM2 digital library keeps all previous XML files available in the backend, but only makes the last one (and the derivative HTML, PDF and plain text files) available to the user.
To make the versioning of the ILC4CLARIN repository coherent with ALIM's, we decide to remove the demo URL from the old version(s). In this way, the users access the last version of the document from the repository and, if they nevertheless need previous data, can contact ALIM and request for the previous data.

[13] `https://wiki.lyrasis.org/display/DSDOC5x/Importing+and+Exporting+Items+via+Simple+Archive+Format`
[14] `https://www.urlencoder.io/learn/`

5 Concluding remarks

The DFCLAM committed itself to offering data and free online access to some digital archives of literary and historical texts: one of them is ALIM the largest digital library of the Italian Latinity including both literary and documentary texts, encoded in XML TEI from philologically checked printed editions or firstly edited from manuscripts, produced in Italy during the Middle Ages. Strategies for importing the metadata of ALIM in the CLARIN-ILC repository through a shared TEI header are under study, as well as procedures for delivering dedicated tools for textual and linguistic analysis through the CLARIN channels. This would allow meta-queries and cross-queries on semantic items which could connect Latin and modern European languages derived from Latin and allow to develop semantic trees and networks of lexical derivations at the very heart of the European shared lexicon.

ALIM complements the Latin resources in CLARIN by providing access to a large corpus of medieval literary and documentary Latin texts with granular curated metadata. On the other hand the VLO makes the resources produced and described in the ILC4CLARIN repository, including ALIM metadata, available to a wider audience in the SSH community, while the CMDI model ensures high quality metadata curation. Also, CLARIN offers ALIM the possibility to use technology and text analysis tools available at CLARIN data centers to deal with multilingual data. For example, Weblicht allows to combine web services so as to handle and exploit textual data.

References

[Alessio2003] Gian Carlo Alessio. 2003. Il progetto alim (archivio della latinità italiana del medioevo). In Francesco Santi, editor, *In Biblioteche elettriche. Letture in Internet: una risorsa per la ricerca e per la didattica*, volume 1, pages 73–81. SISMEL - Edizioni del Galluzzo.

[Buzzoni and Rosselli Del Turco2016] Marina Buzzoni and Roberto Rosselli Del Turco. 2016. Evolution or revolution? digital philology and medieval texts: History of the discipline and a survey of some italian projects. In *Mittelalterphilologien heute. Eine Standortbestimmung. Band 1: Die germanischen Philologien*, pages 265–294. Königshausen und Neumann.

[D'Angelo and Monella2019] Edoardo D'Angelo and Paolo Monella. 2019. ALIM (Archivio della Latinità Medievale dItalia). Storia, attualità, prospettive di una banca-dati di testi mediolatini. In Roberto Gamberini, Paolo Canettieri, Giovanna Santini, and Rosella Tinaburri, editors, *La Filologia Medievale. Comparatistica, critica del testo e attualità. Atti del Convegno (Viterbo, 26-28 settembre 2018)*, volume 3 of *Filologia Classica e Medievale*. L'Erma Di Bretschneider.

[Ferrarini2017] Edoardo Ferrarini. 2017. ALIM ieri e oggi. *Umanistica Digitale*, 1:7–17.

[Nicolas et al.2018] Lionel Nicolas, Alexander König, Monica Monachini, Riccardo Del Gratta, Silvia Calamai, Andrea Abel, Alessandro Enea, Francesca Biliotti, Valeria Quochi, and Francesco Vincenzo Stella. 2018. CLARIN-IT: State of Affairs, Challenges and Opportunities. In *Selected papers from the CLARIN Annual Conference 2017, Budapest, 18-20 September 2017*, Linköping electronic conference proceedings (Print), pages 1–14.