CLARIN

Selected papers from the **CLARIN Annual Conference 2020** Virtual edition

Selected Papers from the

CLARIN Annual Conference 2020

Virtual Event, 2020, 5-7 October

edited by Costanza Navarretta and Maria Eskevich

Linköping Electronic Conference Proceedings 180
eISSN 1650-3740 (Online) • ISSN 1650-3686 (Print) 2021
ISBN 978-91-7929-609-4

Introduction

Franciska de Jong
Executive Director CLARIN ERIC
Universiteit Utrecht, The Netherlands
f.m.g.dejong@uu.nl

Costanza Navarretta
Programme Committee Chair
University of Copenhagen
Copenhagen, Denmark
costanza@hum.ku.dk

This volume presents the highlights of the 9^{th} CLARIN Annual Conference 2020. The conference was held in the virtual format on 5^{th} —7^{th} October 2020 because of the COVID-19 pandemics. CLARIN, the Common Language Resources and Technology Infrastructure, is a virtual platform for everyone interested in language. CLARIN offers access to language resources, technology, and knowledge, and enables cross-country collaboration among academia, industry, policy-makers, cultural institutions, and the general public. Researchers, students, and citizens are offered access to digital language resources and technology services to deploy, connect, analyse and sustain such resources. In line with the Open Science agenda, CLARIN enables scholars from the Social Sciences and Humanities (SSH) and beyond to engage in and contribute to cutting-edge, data-driven research driven by language data.

The infrastructure is run by CLARIN ERIC[1], a consortium of participating countries and institutes that since it was established in 2012 has grown ins size considerably. Currently there are 21 member countries, 3 observers, and more than 100 associated research institutions who are all encouraged and supported to be represented at the annual conference which is meant to be a central event for CLARIN community and which is one of the crucial instruments for CLARIN to function as a knowledge hub. At the conference, consortia from all participating countries and the various communities of use meet, in order to exchange ideas, experiences and best practices in using the CLARIN infrastructure. The conference covers a wide range of topics, including the design, construction and operation of the CLARIN infrastructure, the data, tools and services that are or could be on offer, its actual use by researchers, its relation to other infrastructures and projects, and the CLARIN Knowledge Infrastructure. The aim is to attract researchers from all the various SSH fields who work with language materials, i.e. the people who are the raison d'être for CLARIN. Early in 2020 a call[2] was issued for which 40 abstracts were submitted. The authors of the submissions to the main conference session represented 19 countries, all of them from CLARIN ERIC countries with the exception of one paper from Spain. A few papers were written in cooperation by authors from different countries and institutions, but the number of the cross-country submissions is lower than in the previous conference editions. This can be due to the fact that CLARIN members, as the rest of the world, have not been able to meet each other face two face in most of 2020. Moreover, we did not receive contributions outside Europe, which again could be a negative effect of the pandemic restrictions.

All submissions were reviewed anonymously by three reviewers (PC members and reviewers invited by PC members). Out of the 40 submitted abstracts 36 submissions were accepted for presentation at the conference (acceptance rate 0.9). The submissions were grouped in the following subjects:

- Annotation and Visualization Tools
- Data Curation, Archives and Libraries
- Metadata and Legal Aspects
- Research Cases
- Repositories and Workflows
- Resources and Knowledge Centres for Language and AI Research

[1] http://www.clarin.eu
[2] https://www.clarin.eu/content/call-abstracts-clarin-annual-conference-2020

Franciska de Jong and Costanza Navarretta 2021. Introduction. *Selected papers from the CLARIN Annual Conference 2020*. Linköping Electronic Conference Proceedings 180: 180 i–iv.

The accepted contributions were published in the online Proceedings of the Conference[3].

Following the well received student poster session that was part of the programme of the 2018-2019 editions of the CLARIN Annual Conference, a PhD-session was organised with 7 presentations by PhD-students. One of the PhD-presenters represented a non-CLARIN country. The abstracts of the student presentations were published in the online CLARIN 2020 Book of Abstracts[4].

The 2020 edition of the CLARIN Annual Conference was shaped as an online event. The virtual format enabled us to share quality content with almost 500 registered participants, including attendants of previous editions as well as newbies with an interest in getting familiarised with what CLARIN is about. The conference programme contained both traditional conference elements, and novel items better suited for the virtual set-up:

- **Invited talk** by Dr. Antske Fokkens (Faculty of Humanities, Vrije Universiteit Amsterdam)g ave a talk entitled "Language Technology & Hypothesis Testing". In this talk, she has highlighted that both the quality and accessibility of language technology has drastically increased over the last decade. Generic language models and deep learning have led to impressive results and both models and code for creating and using them is often made available. As such, an increase of these technologies are seen to be used in industry and various research disciplines outside of computational linguistics. Despite sometimes impressive results, however, currently developed technologies are still far from perfect and much is still unknown about how well those models work for specific use cases. Eventually, she has argued for the importance of going back to the foundations and ground research in hypotheses, both for studying language technology itself as well as for applying it in other research domains.

- **Panel on Artificial Intelligence, Language Data and Research Infrastructures** moderated by Ben Verhoeven with the following experts:

  - Prof. Jan Hajic, full professor of Computational Linguistics and the deputy head of the Institute of Formal and Applied Linguistics at the School of Computer Science, Charles University in Prague;
  - Dr Vukosi Marivate, ABSA UP Chair of Data Science at the University of Pretoria;
  - Prof. Marie-Francine Moens, full professor at the Department of Computer Science at KU Leuven, Belgium; director of the Language Intelligence and Information Retrieval (LIIR) research lab, a member of the Human Computer Interaction group, and head of the Informatics section;
  - Prof. dr. Malvina Nissim, Professor in Computational Linguistics and Society at the University of Groningen, The Netherlands; coordinator of the Computational Linguistics Group of the Center for Language and Cognition Groningen.

- Three **Special Appetizers** during the lunch breaks

  - CLARIN Café: "This is CLARIN. How can we help you?" with the aim to give an overview of CLARIN in a nutshell.
  - Social Networking Lunch
  - Improbotics - Improvised Theatre Show

- **Sessions of accepted conference papers** were organised as moderator-led discussions, and followed by poster-style discussions during which session participants could visit the individual paper authors and engage into discussions.

- During the **CLARIN Student session**, PhD-students presented their work in progress . The aim of the session was to share the next generation of researchers supported by or contributing to the CLARIN infrastructure and enable them to receive feedback on their work from CLARIN experts.

[3] https://office.clarin.eu/v/CE-2020-1738-CLARIN2020$_{C}onferenceProceedings.pdf$

[4] https://www.clarin.eu/content/clarin2020-book-abstracts

- The **CLARIN in the Classroom session** invited university lecturers who had used CLARIN resources, tools or services in their courses to present their experience and suggest future steps that could help facilitate and accelerate the further integration of CLARIN into university curricula. (The slides of both sessions can be found in the conference programme).
- The **CLARIN Bazaar** provided as usual an informal setting for conversations with CLARIN people and a space to showcase ongoing work and exchange ideas.
- Each day was finished a **wrap-up session** that combined both personal highlights of two experts in the field and an illustration by by professional sketch artist.

In addition, on the event page[5] CLARIN published a rich set of materials related to the conference:

- The complete conference programme and most of the slides presented: https://www.clarin.eu/content/programme-clarin-annual-conference-2020
- Recordings of keynote, panel, and CLARIN Café that are available on the CLARIN YouTube channel: link to be added.

After the conference, the authors of the accepted papers and student submissions, as well as participants of the CLARIN in the Classroom session were invited to submit full versions of their papers to be considered for the post-conference proceedings volume. The papers were anonymously reviewed, each by three PC members. We received 27 (including 1 student paper and 1 paper by the group of lecturers) full length submissions, out of which 23 were accepted for this volume. All the main topics addressed at the conference are covered in the papers.

We would like to thank all PC members and reviewers for their efforts in evaluating and re-evaluating the submissions, Maria Eskevich from CLARIN Office for her indispensable support in the process of preparing these proceedings, and our colleagues at the Linköping University Electronic Press, who have ensured that the digital publication of this volume came about smoothly. In order to support the programme chair and the programme committee in the organisation of reviewing and programme planning, a programme subcommittee was established starting from CLARIN 2020. With respect to the establishment of the programme subcommittee, it was decided that the programme chair from the preceding year's conference is one of members in order to ensure continuity from one year's conference to the following one. The members of the 2020 PC subcommittee were Eva Hajičová, Monica Monachini, Kiril Simov, Inguna Skadiņa, and Martin Wynne.

Members of the Programme Committee for the CLARIN Annual Conference 2020:

- Lars Borin, Språkbanken, University of Gothenburg, Sweden
- António Branco, Universidade de Lisboa, Portugal
- Koenraad De Smedt, University of Bergen, Norway
- Tomaž Erjavec, Jožef Stefan Institute, Slovenia
- Eva Hajičová, Charles University Prague, Czech Republic
- Martin Hennelly, South African Centre for Digital Language Resources, South Africa
- Erhard Hinrichs, University of Tübingen, Germany
- Marinos Ioannides, Cyprus University of Technology (CUT), Cyprus
- Nicolas Larrousse, Huma-Num, France

[5] https://www.clarin.eu/event/2020/clarin-annual-conference-2020-virtual-event

- Krister Lindén, University of Helsinki, Finland
- Monica Monachini, Institute of Computational Linguistics “A. Zampolli”, Italy
- Karlheinz Mörth, Institute for Corpus Linguistics and Text Technology, Austrian Academy of Sciences, Austria
- **Costanza Navarretta, University of Copenhagen, Denmark (Chair)**
- Jan Odijk, Utrecht University, The Netherlands
- Maciej Piasecki, Wrocław University of Science and Technology, Poland
- Stelios Piperidis, Institute for Language and Speech Processing (ILSP), Athena Research Center, Greece
- Eirikur Rögnvaldsson, University of Iceland, Iceland
- Kiril Simov, IICT, Bulgarian Academy of Sciences, Bulgaria
- Inguna Skadiņa, University of Latvia, Latvia
- Marko Tadič , University of Zagreb, Croatia
- Jurgita Vaičenonienė, Vytautas Magnus University, Lithuania
- Tamás Váradi, Research Institute for Linguistics, Hungarian Academy of Sciences, Hungary
- Kadri Vider, University of Tartu, Estonia
- Martin Wynne, University of Oxford, United Kingdom

Additional reviewers of this volume:

- Iulianna van der Lek-Ciudin, CLARIN ERIC, The Netherlands
- Riccardo Del Gratta, ILC “A. Zampolli” CNR Pisa, Italy

Contents

"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, Sapienza Università
di Roma, Italy
paolo.monella
@uniroma1.it

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

Abstract

This paper aims at showing how integrating the Archive of the Italian Latinity of the Middle Ages (ALIM) into the ILC4CLARIN repository can provide mutual benefits. Making ALIM available to a large community of scholars and researchers, on the one side, represents the first step to reduce the lack of resources for Medieval Latin in CLARIN and, on the other side, constitutes an unprecedented contribution to not only linguistic investigations, but also to the studies of the culture and science at the basis of the Western European society. The paper describes the adopted approach aiming to keep intact the structure of the archive and its metadata, which are both accurately mirrored into the ILC4CLARIN repository in order to maintain existing access practices of the users. This structure can be found in exactly the same state within the CLARIN VLO. Finally, the paper illustrates the advantages of experimenting with some ALIM data, once introduced within the CLARIN Language Resource Switchboard service: first results are shown from the analysis of some texts with the UDPipe tool suite and the distant reading tool Voyant.

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 study not only of Latin, but also of the culture and science at the basis of the Western European society. For several centuries, in fact, Latin represented the only language in which many of the major creations of thought, science, and literature of the Middle Ages were expressed. Even when national languages imposed themselves in written form, Latin never lost its role and prestige as a transnational language – until the end of the Middle Ages and beyond.

Federico Boschetti, Riccardo Del Gratta, Monica Monachini, Marina Buzzoni, Paolo Monella and Roberto Rosselli Del Turco 2021. "Tea for two": the Archive of the Italian Latinity of the Middle Ages meets the CLARIN Infrastructure. *Selected papers from the CLARIN Annual Conference 2020.* Linköping Electronic Conference Proceedings 180: 180 37–46.

The general aim of this paper is to place ALIM within the framework of CLARIN-IT and CLARIN at large. Section 2 shows how ALIM may contribute to fill an important gap in the availability of literary and historical sources for Latin: 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 internal structure and the metadata of the Archive are presented in Section 3, while the strategy for the integration of ALIM into the ILC4CLARIN repository is discussed in Section 4. Finally, ALIM's contribution in strengthening and widening the research directions in CLARIN-IT and its advantages for the CLARIN community are presented.

2 Latin resources in CLARIN-IT

The Italian CLARIN (CLARIN-IT) consortium[1] (Nicolas et al., 2018) has a strong interest in the field of Digital Classics, which is still affected by a shortage or restricted availability of language resources for historical languages such as Ancient Greek and Latin. To this end, the consortium aims to make some of the existing digital resources for Ancient Greek and Latin available through its national repository – ILC4CLARIN (Pisa).

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 ILC4CLARIN data center mostly concerns the study of methods and the development of services to offer online secure access to some digital archives of literary and historical texts. ALIM, currently hosted by the University of Siena[3], is the largest digital library of the Italian Latinity, including both literary and documentary sources.

Evidence shows that the CLARIN data centers do not offer resources such as ALIM. The faceted-search functionality of the Virtual Language Observatory (VLO), performed combining *Latin text resource* and *Middle Ages*, returns 53 records, while 124 records 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 the EUROPEANA platform[4].

ALIM represents the first step to reduce the lack of resources for Medieval Latin in CLARIN-IT and eventually in CLARIN. On the one hand, ALIM will offer high-quality data since: (i) the resources are curated by domain experts; (ii) a strong organization is dedicated to maintenance; (iii) the resources cover a broad historical period; (iv) the resources are TEI encoded. On the other hand, the Language Resource Switchboard and the Weblicht workflow engine will use the texts provided by ALIM to produce both engaging visualizations and interesting linguistic analyses.

3 ALIM: history, goals, structure

ALIM is an archive of medieval Latin texts composed in the Italian area between the 8th and 15th centuries. 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[5].

[1] The composition of the Italian Consortium is available at `http://clarin-it.it/en/content/consortium`.

[2] Prof. Francesco Vincenzo Stella.

[3] `http://alim.unisi.it/il-progetto/`

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

[5] 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).

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[6] and its texts were annotated with 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 on the servers of the University of Siena[7]; (2) re-encoding text markup and metadata in TEI XML P5.

Task 1 was realised in collaboration with the external IT company Net7[8] 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" (Ferrarini, Monella and Rosselli Del Turco) to gradually improve the level of formalisation and the granularity of text markup and metadata.

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[9], 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 the ALIM2 TEI-XML files for literary texts deriving from the initial export from ALIM1 (labeled as "encoding level ALIM2_0"), much metadata was still 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. Ctibor's activity was brought forth in the framework of a collaboration agreement between ALIM and the *Corpus Corporum*[10], the largest full-text repository for Latin (163 M words). The current policy of ALIM requires that all new texts included in the archive must be encoded at "level ALIM2_2": this includes markup of work titles, quotes, speeches, persons, or place names.

The archive also includes born-digital scholarly editions directly based on handwritten medieval witnesses, whose encoding level is labeled as ALIM2_3[11].

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[12]. This structure is important for, at least, two reasons. The first underlying motivation for this partially conservative choice was that this decision would provide philologists, linguists, and historians with a user experience on the VLO consistent with the navigation in the original ALIM environment. The second reason is directly con-

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

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

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

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

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

[11] 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/`

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

nected with the VLO. In section 2, we briefly mentioned the faceted-search of the VLO. One of such facets is the collection (in the original 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"[13].

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 differs 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.

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[14] 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.
More in detail: (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 different metadata files combine to create the descriptive items in the repository. 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. It is important, here, to notice 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*, that is to say where the resource stays in the ALIM digital library.

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 function 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_doc`

[13] 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/xmlui/handle/000-c0-111/130`.

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

umentaria/7381#doc_27. Unfortunately, '#' is a reserved character[15] 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.

As an example, Figure 1 shows the different elements in the descriptive item mapped to their sources in the TEI header.

Figure 1: Repository item and its sources in the TEI Headers.

Before concluding this section, let us provide some information on the licence of the data. The original data, contained in the ALIM digital library, are released under the Creative Commons - Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND4.0). To stay compliant with the original license, the loading procedures adds this information to the descriptive items, as reported in Figure 2.

dc.rights.uri	http://creativecommons.org/licenses/by-nc-nd/4.0/
dc.rights.label	PUB

Figure 2: Additional metadata inform about licences and IPR.

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 file 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.

[15] https://www.urlencoder.io/learn/

To make the versioning of the ILC4CLARIN repository consistent with that of ALIM, we decided to remove the *demo URL* from the old versions. In this way, users access the latest version of the document from the repository and, if they still need older data, they can contact ALIM and request them.

5 CLARIN Services and ALIM

5.1 Available analysis tools

In this section, we show the results of an analysis of an ALIM sample text with tools made available through the CLARIN infrastructure.
The Language Resource Switchboard (LRS)[16] (Zinn, 2018) has been used to connect the input data with suitable and available tools. Figure 3 below shows the suggested tools for the input file "Historia Langobardorum"[17], which consists of about 38000 words.

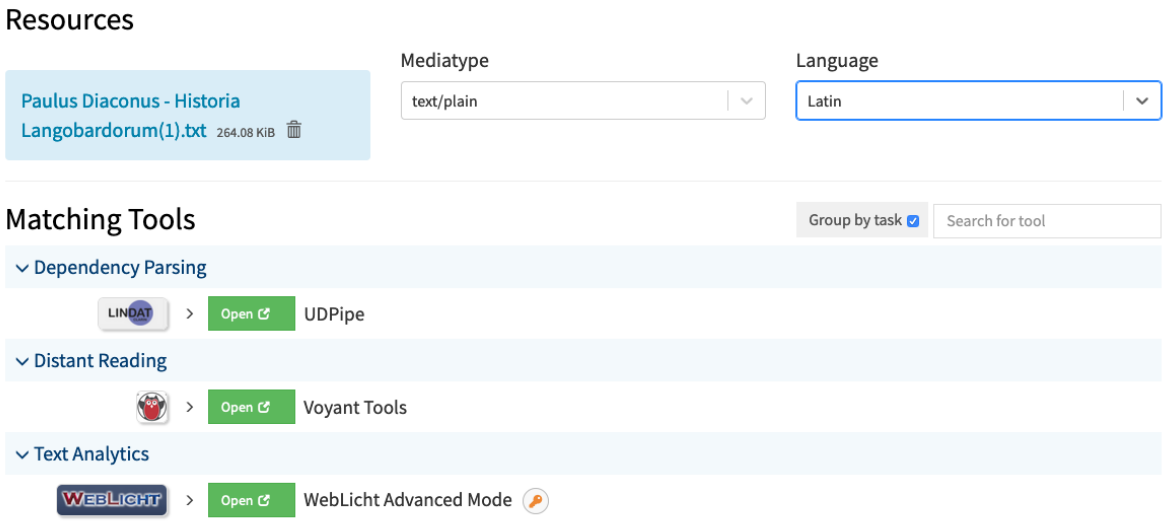

Figure 3: "Historia Langobardorum" and connected tools.

The LRS lists a distant reading tool, Voyant[18] and a dependency parsing tool, UDPipe (Straka and Straková, 2020; Straka and Strakova, 2017).

Figure 4 displays the *Cirrus* and *TermsBerry*, while Figure 5 provides some textual statistics on the examined text[19].

[16] https://switchboard.clarin.eu/
[17] http://hdl.handle.net/20.500.11752/OPEN-152.
[18] https://voyant-tools.org/
[19] We invite the interested readers to replicate the experiment by providing the URL http://alim-admin.unisi.it/download_txt?id=201 to the LRS and eventually use *Voyant tools*.

Figure 4: Cirrus and TermsBerry

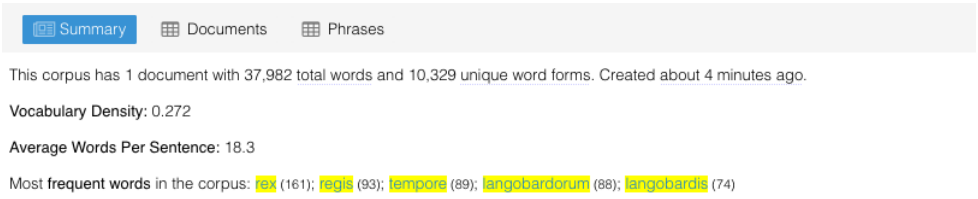

Figure 5: Statistics of the input texts.

The dependency parser tool, UDPipe, is both available as a service[20], from LINDAT/CLARIAH-CZ[21], and as an integrated service into the WebLicht workflow engine[22] (Hinrichs et al., 2010).
The following figures show the analysis of the sentence "Inter haec moritur Godehoc, cui successit Claffo, filius suus"[23], extracted from the same text, performed using the UDPipe parser from LINDAT/CLARIAH-CZ. The model used is the *latin-llct-ud-2.6-200830*.

[20] `https://lindat.mff.cuni.cz/services/udpipe/info.php`.
[21] `https://lindat.mff.cuni.cz`.
[22] Again here, we recommend the interested readers at replicating the experiment, focusing on the different access modalities.
[23] With more complex sentences the results might not be completely correct, both in terms of lemmatization and syntactic trees.

text = Inter haec moritur Godehoc, cui successit Claffo, filius suus

1	Inter	inter	ADJ	AAXXX----1A----	Degree=Pos|Foreign=Yes|Polarity=Pos	0	root	_	TokenRange=0:5
2	haec	haec	NOUN	AAXXX----1A----	Degree=Pos|Foreign=Yes|Polarity=Pos	1	flat:foreign	_	TokenRange=6:10
3	moritur	moritur	NOUN	NNMS1-----A----	Animacy=Anim|Case=Nom|Gender=Masc|Number=Sing|Polarity=Pos	1	flat:foreign	_	TokenRange=11:18
4	Godehoc	Godehoc	PROPN	NNMS1-----A----	Animacy=Anim|Case=Nom|Gender=Masc|NameType=Giv|Number=Sing|Polarity=Pos	1	flat:foreign	_	SpaceAfter=No|TokenRange=19:26
5	,	,	PUNCT	Z:----------	_	7	punct	_	TokenRange=26:27
6	cui	cui	PRON	PQ--4----------	Animacy=Inan|Case=Acc|PronType=Int,Rel	7	obl:arg	_	TokenRange=28:31
7	successit	successit	VERB	VpYS---XR-AA---	Aspect=Perf|Gender=Masc|Number=Sing|Polarity=Pos|Tense=Past|VerbForm=Part|Voice=Act	3	acl:relcl	_	TokenRange=32:41
8	Claffo	Claffo	PROPN	NNMS1-----A----	Animacy=Anim|Case=Nom|Gender=Masc|NameType=Giv|Number=Sing|Polarity=Pos	7	nsubj	_	SpaceAfter=No|TokenRange=42:48
9	,	,	PUNCT	Z:----------	_	10	punct	_	TokenRange=48:49
10	filius	filius	NOUN	NNMS1-----A----	Animacy=Anim|Case=Nom|Gender=Masc|Number=Sing|Polarity=Pos	8	appos	_	TokenRange=50:56
11	suus	suus	NOUN	NNIS1-----A----	Animacy=Inan|Case=Nom|Gender=Masc|Number=Sing|Polarity=Pos	10	nmod	_	SpaceAfter=No|TokenRange=57:61

Figure 6: A sample UDPipe table.

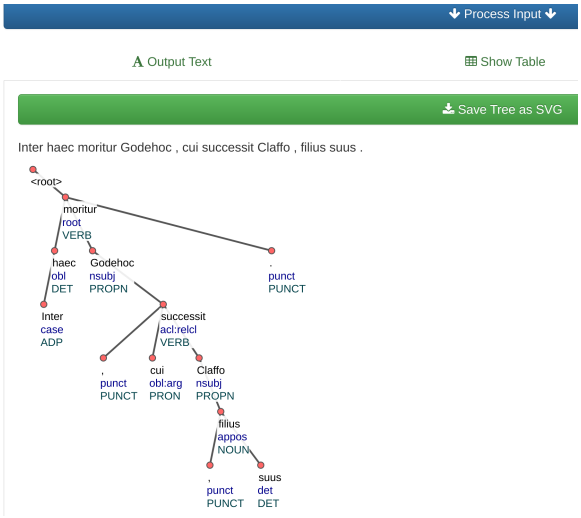

Figure 7: A sample UDPipe tree.

A few words on the use of WebLicht for the ALIM texts. As a workflow engine, WebLicht can run as many tools as needed. Figure 8 reports the analysis chain which can be run on the same sentence. The chain consists of three modules: a tokenizer, a tagger, and a parser. The tokenized text can be further used in the Federated Content Search (Stehouwer et al., 2012), cf. Section 6.

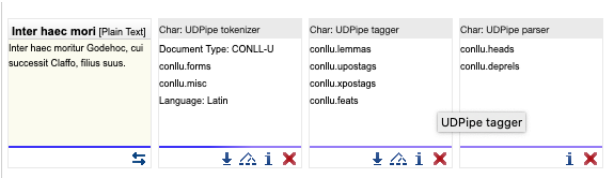

Figure 8: WebLicht analysis chain.

5.2 Are similar resources available?

The "Historia Langobardorum" was already available from the VLO as a digital facsimile of the medieval manuscript provided by e-codices, the Virtual Manuscript Library of Switzerland (`https:`

//www.e-codices.ch/en/csg/0635/1/0/). Even if the direct source of the text of "Historia Langobardorum" in ALIM is the digitization of the nineteenth-century edition by L. Bethman - G. Waitz[24], the manuscript available from the VLO is one of its witnesses. Indeed, the increased use of the IIIF (www.iiif.io) framework to publish digital facsimiles of medieval manuscripts means that for many of the ALIM texts there are already one or more digitized witnesses freely available on the Web. Not only that, this opens the way for a future connection of an ALIM text to those facsimiles: for the most advanced ALIM critical editions, those belonging to level 3 of the encoding, i.e. a full use of the elements of the Critical Apparatus TEI module (cf. Section 3.1), this could also mean generating automatically the corresponding witness and linking it to the digitized images of the manuscript. s

6 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 Archive of the Italian Latinity of the Middle Ages), the largest digital library of the Italian Latinity including both literary and documentary texts, encoded in TEI XML from philologically checked printed editions or published directly from manuscripts produced in Italy during the Middle Ages, in new born-digital scholarly editions. Strategies for importing the metadata of ALIM in the ILC4CLARIN 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 vocabulary.

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, participating in CLARIN provides ALIM with a valuable opportunity in terms of sustainability, long term preservation, persistent identification, format migration, and visibility for the ALIM research outputs. 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, while the Language Resource Switchboard can connect the ALIM texts to visualization tools such as Voyant. A further reciprocal advantage is that CLARIN contributed in enhancing the ALIM strategies on Open Access and open source policies by supporting ALIM in planning the actions necessary to provide FAIR (Findable, Accessible, Interoperable, Reusable) data (de Jong et al., 2018).

Finally, ILC4CLARIN is an endpoint of the Federated Content Search (FCS) [25], a tool to query data distributed across local collections available at the various CLARIN centres. At the time of writing, the only CLARIN collection including Latin texts (lat_wikipedia_2012_100K, about 1.5M tokens) is available at the FCS aggregator from the Automatische Sprachverarbeitung - Universität of Leipzig. The addition of further sources to the aggregator will be of fundamental importance to increase the number of high-quality Latin texts available through the Federated Content Search.

[24] MGH SS rer. Lang., Hannover 1878, pp. 45-18

[25] https://contentsearch.clarin.eu/.

References

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.

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.

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.

Franciska de Jong, Bente Maegaard, Koenraad De Smedt, Darja Fišer, and Dieter Van Uytvanck. 2018. CLARIN: Towards FAIR and responsible data science using language resources. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

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

Erhard Hinrichs, Marie Hinrichs, and Thomas Zastrow. 2010. WebLicht: Web-based LRT services for German. In *Proceedings of the ACL 2010 System Demonstrations*, pages 25–29. Association for Computational Linguistics.

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.

Herman Stehouwer, Matej Durco, Eric Auer, and Daan Broeder. 2012. Federated search: Towards a common search infrastructure. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Milan Straka and Jana Strakova. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. pages 88–99, 01.

Milan Straka and Jana Straková. 2020. Udpipe at evalatin 2020: Contextualized embeddings and treebank embeddings. *arXiv preprint arXiv:2006.03687*.

Claus Zinn. 2018. The language resource switchboard. *Comput. Linguist.*, 44(4):631–639, December.