

Selected papers from the
CLARIN Annual Conference 2023

Leuven, Belgium, 16–18 October 2023

edited by Krister Lindén, Thalassia Kontino and Jyrki Niemi



Front Cover Illustration:

Picture Composition by CLARIN ERIC

Licensed under Creative Commons Attribution 4.0 International:

<https://creativecommons.org/licenses/by/4.0/>

Linköping Electronic Conference Proceedings
ISSN 1650-3740 (digital) • ISSN 1650-3686 (print)
ISBN 978-91-8075-740-9 (PDF)

210
2023

Introduction

Krister Lindén

Program Committee Chair
University of Helsinki,
Finland
krister.linden@helsinki.fi

Darja Fišer

Executive Director of CLARIN ERIC
Institute of Contemporary History,
Slovenia
darja.fiser@clarin.eu

OMISSIS

¹<http://www.clarin.eu>

²<https://www.clarin.eu/content/call-abstracts-clarin-annual-conference-2023>

The authors of the accepted papers, student submissions, participants of the workshop ‘Using CLARIN in Training and Education’ as well as CLARIN-funded projects were invited to submit papers for the post-conference proceedings. Out of the conference contributions, 17 substantially extended contributions were accepted as Selected papers from the CLARIN Annual Conference 2023.

We would like to thank all PC members and reviewers for their efforts in evaluating and re-evaluating the submissions, Thalassia Kontino from the CLARIN Office for her indispensable support in the process of preparing these proceedings, and our colleagues at the Linköping University Electronic Press, who 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. To ensure continuity, the programme chair from the preceding year’s conference was a member of the committee. The members of the 2023 programme subcommittee were Krister Lindén (chair), Tomaž Erjavec, Monica Monachini, Maciej Piasecki and Vincent Vandeghinste.

Members of the Programme Committee for the CLARIN Annual Conference 2023:

- **Krister Lindén, University of Helsinki, Finland (Chair)**
- Starkaður Barkarson, Árni Magnússon Institute for Icelandic Studies, Iceland
- Lars Borin, University of Gothenburg, Sweden
- António Branco, Universidade de Lisboa, Portugal
- Koenraad De Smedt, University of Bergen, Norway
- Tomaž Erjavec, Jožef Stefan Institute, Slovenia
- Cristina Grisot, Switzerland
- Eva Hajičová, Charles University Prague, Czechia
- Marinos Ioannides, Cyprus University of Technology, Cyprus
- Monica Monachini, Institute of Computational Linguistics “A. Zampolli”, Italy
- Costanza Navarretta, University of Copenhagen, Denmark
- Maciej Piasecki, Wrocław University of Science and Technology, Poland
- Stelios Piperidis, Institute for Language and Speech Processing (ILSP), Athena Research Center, Greece
- German Rigau, Basque Center for Language Technology, Spain
- Gijsbert Rutten, Leiden University, The Netherlands
- Kiril Simov, IICT, Bulgarian Academy of Sciences, Bulgaria
- Inguna Skadiņa, University of Latvia, Latvia
- Marko Tadić, University of Zagreb, Croatia
- Jurgita Vaičenonienė, Vytautas Magnus University, Lithuania
- Vincent Vandeghinste, Instituut voor de Nederlandse Taal (Dutch Language Institute), the Netherlands & KU Leuven, Belgium

Domain-Specific Languages for Epigraphy: the Case of ItAnt

Federico Boschetti

CNR-ILC

URT Venezia, Italy

`federico.boschetti@ilc.cnr.it`

Luca Rigobianco

Dipartimento di Studi Umanistici

Università Ca' Foscari Venezia

Venezia, Italy

`luca.rigobianco@unive.it`

Valeria Quochi

CNR-ILC

Pisa, Italy

`valeria.quochi@ilc.cnr.it`

Abstract

This contribution illustrates how the definition of a Domain-Specific Language can support the activities of epigraphists and historical linguists. It presents and discusses a method and technological solution, based on Domain-Specific Languages, for facilitating scholars in digitally representing the available knowledge of archaic languages and cultures. This is achieved by increasing the human readability of the encoded data without sacrificing compliance with standard models and formats. The work is framed within the context of an Italian National collaborative research project devoted to the study of the languages and cultures of ancient Italy. The platform developed within this project offers an interesting use case and motivation for experimenting with Domain-Specific Languages for the creation of necessary digital critical editions of the inscriptions relevant for these languages. After explaining the definition process of the DSL grammar, we finally test the applicability of the DSL grammar to five example inscriptions in the Faliscan language.

1 Introduction

The recovery, digitisation, and sharing of knowledge relating to ancient fragmentary languages and their cultures are primary concerns within the fields of historical linguistics and digital humanities, posing significant challenges. Fragmentary languages are dead languages attested through a highly restricted corpus of surviving texts. Such a corpus is limited due to socio-cultural choices on what to write as well as the randomness of the documentary findings. Due to these restrictions, the knowledge that can be derived is necessarily partial and sometimes uncertain, both in terms of grammar and lexicon, and with regard to language variation over time and space, along the social ladder, and according to the communicative situation. Such an incomplete, uncertain, and quantitatively scarce written evidence hampers the use of state-of-the-art AI or machine learning techniques and requires an adaptation of existing language technology. This can only be achieved through the collaboration between historical linguists and language technologists.

The very first and fundamental stage in this direction is the creation of robust, machine-actionable digital scholarly editions of the inscriptions and their linguistic content, a task what is by all means non-trivial. Recently, the ILA project (Sarullo, 2016) has taken a first step towards adapting the XML-TEI/EpiDoc standard model to an epigraphically attested fragmentary language such as archaic Latin (7th-5th century BC). Additionally, the i.Sicily digital corpus (Prag & Chartrand, 2019) deserves to be mentioned, as it collects texts from ancient Sicily dating from the 7th century BC to the 7th century AD, including fragmentary languages such as Sikel and Elymian.

In general, despite the considerable effort required, the challenge of adequately treating these languages digitally must be faced. This is necessary in order to preserve their documentation and knowledge and make them widely accessible. However, digitising scholarly editions proves to be a time-consuming and unfriendly task for many scholars. This contribution tackles this issue and introduces a method and technological solution based on Domain-Specific Languages (DSLs hereafter) to facilitate such tasks.

The paper is organised as follows: Section 2 describes the project that motivated this work and the online platform that will consume the produced critical editions, making them available to scholars for

creating interlinked lexical resources. Here the connection to the CLARIN infrastructure is also made explicit. Section 3 gives a brief introduction to DSLs and their advantages in Digital Humanities (DH) contexts before delving into the specific grammar designed for the languages of ancient Italy. Section 4 then demonstrates the applicability of the DSL grammar to five example inscriptions in the Faliscan language. Finally, Section 5, wraps up and suggests possible future directions.

2 The context: the ItAnt project

The project *Languages and Cultures of Ancient Italy. Historical Linguistics and Digital Models* (ItAnt hereafter) is a collaborative initiative funded by the Italian Ministry of University and Research. It aims to investigate the languages of ancient Italy by combining the methods of historical linguistics with digital technologies specifically designed to create a set of interconnected resources, particularly critical digital editions of inscriptions, lexica, and bibliographies¹.

With the sole exception of Roman Latin, the languages of ancient Italy (8th century BC-1st century AD) are fragmentary languages. Their evidence consists almost exclusively of epigraphic texts, which often present problems relating to the reading, segmentation into words, linguistic analysis, and interpretation. Therefore, one of the key challenges of the ItAnt project is to adapt existing digital methods and tools, practices, and methodologies of digital epigraphy and computational lexicography to the highly fragmentary nature of such documentation. Among the languages of ancient Italy, the project focuses on Oscan, Faliscan, Venetic, and Cisalpine Celtic, chosen as representative due to the quantitative and qualitative differences in their documentation and to their belonging to linguistic (sub)groups which are diverse as regards their genetic classification (Pocchetti, 2017).

The main objectives of the project are thus to create and interlink a digital archive of critical editions of inscriptions, a multilingual computational lexicon, and a bibliographic dataset of relevant cited works in FRBRoo²/LRMoo³. Within the digital archive, the inscriptions are encoded in XML following the XML-TEI/EpiDoc model and schema⁴. Furthermore, these new editions of the inscriptions will be enriched with metadata defined in shared common vocabularies, enabling accurate semantic description of them as both linguistic and material objects. The DSL solution described in this paper is responsible for producing these enriched EpiDoc editions of the inscriptions, which will then be ingested by the DigItAnt platform for linking to the other resources mentioned above (see section 2.1 below for details on the platform).

2.1 The DigItAnt platform

Together with the production and publication of datasets for the four languages in focus, one of the main outcomes of ItAnt is a web platform for creating and then exploring the interlinked ecosystem of resources mentioned above: LOD-compliant lexica, critical editions of inscriptions, citations and bibliographic references, as well as other external available salient vocabularies and lexicons.

Assuming that more intuitive disciplinary editing tools can simplify the work of philologists and historical linguists in managing lexical and linguistic knowledge about ancient languages, the DigItAnt platform is designed to assist scholars in encoding lexical information of ancient languages and linking it to other relevant (re-)sources according to semantic web principles. Lexicon creation lies at the heart of the editing platform, which further enables scholars to enrich lexica with actionable links to related inscription, cited bibliographic items, and to other external datasets. Particularly central to the platform is the linking of lexical and morphological forms to their attestations in the texts encoded in XML-TEI/EpiDoc digital scholarly editions of relevant inscriptions⁵.

Digital critical editions of inscriptions, while vital for scholars to consult, play a supporting role in the current version of the editing platform. They are considered instrumental because the platform, in

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

¹For more information see the project website <https://www.prin-italia-antica.unifi.it/>

²<https://cidoc-crm.org/frbroo/>

³LRMoo is the new ontological bibliographic model developed from FRBRoo. A stable version was released only in October 2023.

⁴<http://www.stoa.org/epidoc/gl/latest/>

⁵See Quochi, Bellandi, Mallia, et al. (2022) and Quochi, Bellandi, Khan, et al. (2022) for additional details on the platform.

its current form, offers tools allowing only for the creation and updating of lexicons and the linking of lexical items to external resources, particularly for linking to inscriptions to describe their attestations. As a result, the platform expects inscriptions to be encoded independently in XML according to the XML-TEI/EpiDoc format, the de facto standard for digital epigraphic projects. Consequently, digital editions of inscriptions are considered external datasets, i.e., prepared separately, which the platform can ingest.

Within the ItAnt project, however, we have experimented with the use of a DSL, described in this article, as an alternative to the commonly used Oxygen XML editor. This DSL system offers scholars a lighter and more intuitive way to produce their digital editions. Thus, editions encoded with ItAntDSL, and then converted to EpiDoc XML as described in section 3, can later be ingested by the platform for linking and exploration purposes. In detail, the basic workflow involves historical linguists uploading one or more EpiDoc XML documents to the platform and operate to link specific text segments to either existing lexical items or newly created ones (as exemplified in Fig. 1, see also Quochi, Bellandi, Khan, et al. (2022)).

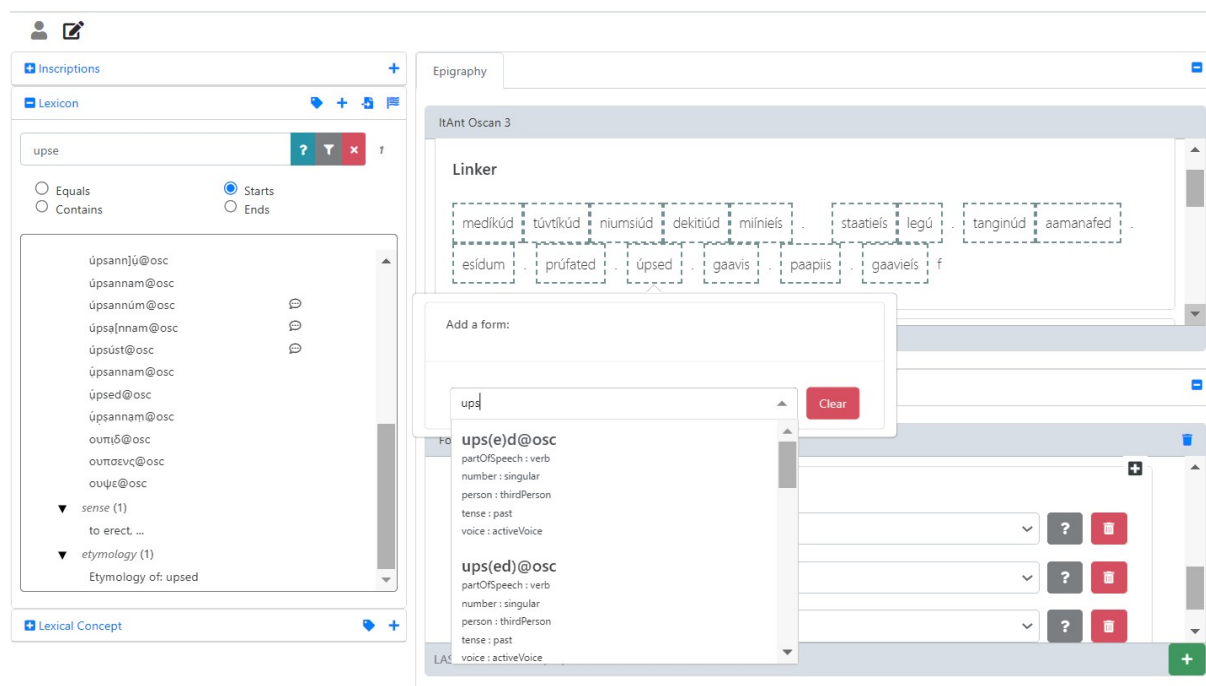


Figure 1: The DigItAnt editor: text - lexicon linking

Thanks to the EpiDoc XML encoding, a visualisation of contextual information as well as of the text according to the Leiden conventions⁶ is also possible, both in the editor (EpiLexO) and in the exploration platform (DigItAnt-search), as shown in Fig. 2⁷.

2.2 The Digital Edition of the Inscriptions

As mentioned above, the project envisages the inscriptions being encoded according to the XML-TEI/EpiDoc schema. Such a schema is the result of an international effort aimed at customising the Text Encoding Initiative's standard for the representation of epigraphic documents according to the Leiden Conventions. In particular, XML-TEI/EpiDoc provides markup for the text (edition, apparatus, translation, commentary, bibliography) as well as the materiality and history of the object on which the text appears (repository, support, layout, hand, place and date of origin, provenance).

⁶The Leiden conventions refer to the system of diacritics used in the scholarly edition of ancient texts, including epigraphic texts; in this regard, refer to Krummrey and Panciera (1980). Since the actual use of these conventions shows a certain variability, for the DSL we have specifically relied on the Leiden+ system, see https://papyri.info/docs/leiden_plus.

⁷The DigItAnt platform prototype is ready and already in use within ItAnt. It's an open-source code available at <https://github.com/DigItAnt>. It is currently maintained and continues to be improved with new functionalities.

The screenshot shows the 'ItAnt Oscan 3' interface. At the top, there are navigation tabs: Home, Inscriptions (selected), Lexicon, Bibliography, Concordances, EpiLexO Search, and Advanced Search. On the left, there is a sidebar with options: Show XML, Print/Save PDF, and Show/Hide restorations. Below this is a map showing the location of the inscription. The main content area is titled 'Text' and contains two lines of Oscan text with red highlighting under certain words. Below the text is a 'Translation' section with a partial Italian translation. A pop-up window for the word 'upsed@osc' is open, showing its type as 'lexicalForm', its morphology as 'partOfSpeech verb' and 'voice activeVoice', and a partial translation: 'Deditius son of son of Status d (this work) in); Gavius Papius son (...)?'.

Figure 2: Text-lexicon data mesh-up in the Explorationplatform

Furthermore, thanks to the extensibility of XML and the versatility of XML-TEI/EpiDoc, ItAnt has proposed solutions for managing specific issues arising from the nature of the languages of ancient Italy as fragmentary languages and their specific epigraphic features (Murano et al., 2023). The customisation has mainly consisted of adding tags to the standard XML-TEI/EpiDoc set. Specifically, within the `<scriptNote>` element, we have opted to specify through `@type` attributes of a `<rs>` element the word division, assuming *'scriptio continua'*, *'punctuation'*, *'blank spaces'*, and *'mixed'* as possible values, as well as the application and simplification of syllabic punctuation for the Venetic inscriptions⁸. Moreover, `<tei:rs>` elements have been added within the `<tei:support>` element to specify the object shape and the possible reuse of the support. A `<rs>` element have also been added within the `<layout>` element to specify if the inscription is opisthographic.

Additionally, a major problem with the XML-TEI/EpiDoc guidelines concerns the theoretical need for a clear distinction between a language and the system(s) used for writing it, since languages and scripts should be indicated together through a single `@ident` attribute of the `<tei:language>` element within the TEI header. For both overcoming such an issue and ensuring interoperability with other digital corpora, we have still followed the guidelines, but we have specified the script(s) regardless of the language(s) through a `<tei:rs>` element within the `<tei:scriptNote>` element, as well as the language(s) regardless of the script(s) through a further `<tei:language>` element. From a linguistic point of view, we have chosen to explicitly mark up the words through `<tei:w>` and `<tei:name>` elements, in order to make it possible to link them to the entries of the computational lexicon. In particular, each word is uniquely identified through an `@xml:id` attribute, whereas the provisioning of lexical information is dealt with in the companion lexicon. The identifier is built using information about the language, the line number, and the position of the word in the line, so as to be transparent and easily readable also by scholars. For example, the value `Fal_6_1_1_w_2` stands for 'second word of the first line of the sixth Faliscan inscription of the ItAnt collection'. The `<tei:name>` elements are further specified through a `@type` attribute as *'praenomen'*, *'gentilicium'*, *'patronymic'*, etc. Furthermore, the use of a `@ref` attribute makes it possible to clearly identify onomastic formulas even in the case of a syntactic break between their components or of a component shared by two or more formulas. The onomastic formulas are then resumed in the commentary (`<tei:div type="commentary">`) through the `<listPerson>` element.

⁸On the peculiar Venetic syllabic punctuation see Marinetti (2020)

With the goal of data integration, ItAnt makes use of widely used vocabularies and gazetteers, in particular *The Art and Architecture Thesaurus* provided by *The Getty Research Institute* is used for specifying object type, material, and writing technique⁹, the EAGLE vocabulary for the type of inscriptions (dedicatory, funerary, etc.)¹⁰, and Pleiades and GeoNames for ancient and modern names respectively¹¹. In addition, Trismegistos IDs are used, when available, to identify the texts¹² and bibliographical records are also linked through a specific library built up by using Zotero¹³.

2.3 Relation with CLARIN

Part of the mission of the ItAnt project is to contribute and integrate data and tools into European Research Infrastructures for the Humanities and Social Sciences, particularly within CLARIN (Common Language Resources and Technology Infrastructure). Since its start, ItAnt has been a project of interest for CLARIN-IT, also due to its potential contribution to the involvement of the community of historical linguists. At the end of the project, the hosting of the platform will transition to the ILC4CLARIN center, where it will be offered as a sustainable open service. Not only will the platform be preserved, but all data and software components will also be deposited and accessible in the long term through CLARIN, for documentation and re-use. The ILC4CLARIN repository¹⁴ already stores copies of LexO-server (Bellandi, 2019), EpiLexO (Mallia et al., 2023) and ItAntDSL (Boschetti & Rigobianco, 2023). At the conclusion of the project, the inscription corpora, lexicons and bibliographies will also be deposited, making them easily discoverable and consumable via CLARIN channels.

Furthermore, due to ItAnt's focus on outputting Linguistic Linked Open Data (LLOD) compatible versions of the data, it will contribute to the development of a CLARIN(-IT) LLOD Platform. In this regard, DigItAnt is a candidate use case for one of the pilot projects to be developed in the context of a recently started large-scale Italian infrastructural initiative, the Humanities and Heritage Italian Open Science Cloud (H2IOSC)¹⁵.

3 Domain-Specific Languages for the encoding of fragmentary archaic languages

3.1 Domain-Specific Languages

Domain-Specific Languages (DSLs) are programming or markup languages created specifically for a certain area of interest. Unlike general-purpose programming languages, which are made to handle a wide variety of programming tasks, DSLs are optimised for a specific field. They aim to provide more expressive power, simplicity, and efficiency for those specific areas. The main benefit of a DSL is its ability to let users describe concepts and actions in ways that closely match the specific abstractions of that domain.

DSLs in the domain of digital epigraphy provide scholars with a set of specialised tools for describing the structure and semantics of inscriptions, enabling precise and detailed digital representations of them. Furthermore, this may enhance the accessibility and dissemination of inscriptions in digital formats. Thanks to DSLs, digital epigraphists can more effectively engage with the textual data, automate repetitive tasks, and focus on the nuanced interpretation and study of the inscriptions.

3.2 How (ItAnt)DSL Facilitates the Encoding

Encoding epigraphic contextual metadata and textual data in XML-TEI/EpiDoc is a complex, error-prone task. Indeed, XML-TEI is quite verbose (because element names, attributes and values must be written in full) and redundant (because opening and closing tags repeat the element names). The percentage

⁹<https://www.getty.edu/research/tools/vocabularies/aat/>. Among other concepts, the AAT taxonomy defines useful terms for describing physical cultural objects such as materials (e.g., pottery, bronze, . . .), object types (e.g., bowl, stele, . . .), and writing techniques (e.g., engraving, inscribing, . . .). Additionally, the iDAI.thesauri provided by the Deutsches Archäologisches Institut (<http://thesauri.dainst.org/de.html>) is used as a supplement with regard to natural supports such as cliffs.

¹⁰<https://www.eagle-network.eu/resources/vocabularies/typeins/>

¹¹<https://pleiades.stoa.org/>; <https://www.geonames.org/>

¹²<https://www.trismegistos.org/tm/>

¹³<https://www.zotero.org/groups/2552746/>

¹⁴<https://ilc4clarin.ilc.cnr.it/>

¹⁵<https://www.h2iosc.cnr.it/home/>

of informative and structural contents is unbalanced. XML-TEI ensures data interchange among software applications and promotes machine actionability and interpretability, but human readability of an encoded document decreases rapidly as complexity increases.

In ItAnt linguistic, philological and prosopographical data are highly entangled. Each word is associated to its part of speech, conjectural integrations to textual gaps (*lacunae*) are recorded, and named entities are identified. These chunks of information often overlap: for instance a lacuna in a line of text may extend between the end of the third token and the beginning of the fourth one, whereas a named entity defined by *praenomen* (partially conjectured), *gentilicium* and *patronymicus* may extend from the fourth to the sixth token.

The problem of overlapping hierarchies in TEI is well-known and many solutions are available, both through manual encoding of stand-off annotations in XML (Spadini & Turska, 2019) and through alternative representations (e.g. in JSON), currently or planned to be convertible in XML-TEI (Neill & Schmidt, 2021). An experimental solution adopted in ItAnt for encoding part of the corpus, is based on a domain-driven approach, which involves the epigraphists to co-design a Domain-Specific Language (Parr, 2009), named ItAntDSL, to encode data and metadata.

The aims of this approach are twofold: a) optimising the encoding process and the encoded documents according to six dimensions (familiarity, transparency, completeness, compactness, consistency, and actionability (Zenzaro et al., 2022) and b) complying with the EpiDoc abstract model. With regard to the above mentioned dimensions, familiarity refers to the maintenance of the scholar's work habits and transparency indicates the level of cognitive effort and/or technical training required of the scholar. Completeness refers to the amount of information which may be expressed, while the ratio between completeness and formalisation is what is meant by compactness. In particular, what occurs more frequently is expected to be encoded with a smaller number of characters. Consistency assesses the coherence in describing the same phenomena in the same way, implying that the representation of the same type of information is unique and therefore unambiguous.

Finally, the ability to extract or deduce information from data is referred to as actionability, which is an intrinsic characteristic in formal languages described by a grammar and commonly accompanied by other components for code processing such as a lexer and a parser. It is evident that a DSL allows for a greater degree of familiarity, transparency, and compactness than an XML encoding. Specifically, once the DSL has been suitably designed by researchers in close contact with experts in the field in question (in our case the fragmentarily attested languages of ancient Italy), it may also be used by scholars who do not know XML nor the XML-TEI/EpiDoc standard, thus drastically reducing the training time necessary to proceed with text encoding. Furthermore, the encoding of contextual metadata (Fig. 3) and textual data (Fig. 4) is very compact. From the user's perspective, this guarantees greater readability and, therefore, the possibility of keeping the text under control, significantly reducing the risk of errors or omissions. In this regard it should also be noted that, although a DSL in itself provides less control over text insertion, the use of an editor may help the scholar by signalling syntactic errors, providing suggestions for their resolution as well as self-completion.

ItAntDSL is defined by a Context-Free Grammar (CFG) available on GitHub¹⁶. The documents encoded in ItAntDSL are then parsed by ANTLR (Parr, 2013), which first converts the Domain-Specific Language into XML with a proprietary schema (XML-ItAnt), based on the production rules of the CFG.

Then, a chain of XQuery scripts and XSLT stylesheets transforms XML-ItAnt documents into XML-TEI/EpiDoc documents. The transformations are not limited to the translation of element names and to structural modifications, but extend to the integration of a) automatically generated IDs; b) default values omitted in ItAntDSL documents; c) expansion of complex structured data encoded in ItAntDSL documents by reference (between quotation marks) and retrieved from the XML documents stored in an eXist-db¹⁷. A sample of the final result is shown in Fig. 5.

As already mentioned, the provision of lexical information is supplied in the companion lexicon and therefore, as far as lexical aspects are concerned, ItAntDSL is limited to explicitly mark up lexical items

¹⁶<https://github.com/CoPhi/itantdsl/>

¹⁷<https://exist-db.org> is a versatile native-XML database management system commonly used in DH projects for managing XML corpora and archives.

```

4 IDENTIFIERS
5 #place: "Schiavi d'Abruzzo (Chieti)"
6 #inst: "in situ (under the tutelage of the Soprintendenza Archeologia, Belle Arti e Paesaggio dell'Abruzzo)"
7 #msName: mosaic from the sanctuary of Schiavi
8 #tm: "TM_170843"
9 #trad: "ST_Sa_2" "ImIt_Teruentum_36"
10
11
12 SUMMARY
13 Inscription recording building and dedication of the paving from temple B of Pietrabbondante sanctuary.
14
15 SUPPORT
16 "temple floor" "tesserae (mosaic components)" #w: 350
17 #notRe-used #very_fragmentary (The inscription is damaged; reading is only possible through photographic material)
18
19 LAYOUT
20 #columns: 1 #writtenLines: 2
21 #exec: "mosaic (opus signinum)" #notOpistograph
22
23 HAND, SCRIPT, AND DECORATION
24 #palaeographicNotes: Letters measure 12 cm in height
25 #characterDimension: 12
26 #alphabet: "Oscan national alphabet"
27 #punctuation
28 |

```

Save

Figure 3: ItAntDSL: metadata

```

46 DIPLOMATIC EDITION
47 #face_a | #text_direction_r_to_l | #sinistrorse
48
49 1 m t ni d!e![.4]ú![.1] [.2] . [.10-12] s!t! legú . tanginúd
50 2 aama!nfed . es!í[.3] . [.6]e!d . ú!psed . g . paapi . g f
51
52
53 ***
54
55 |
56 INTERPRETATIVE EDITION
57 #face_a | #text_direction_r_to_l | #sinistrorse
58
59 1 * m(edíkúd) t(úvtíkúd) ni(umsiúd) d!e![kiti]ú![d] [mi](ínieís) . [10-12] s!t!(aatieís) legú . tanginúd
60 2 * aama!n(a)fed . es!í[dum] . [prúfat]e!d . ú!psed . g(aavis) . paapi(is) . g(aavieís) f()
61
62
63 #line: 1
64 1 m(edíkúd) = #word
65 2 t(úvtíkúd) = #word
66 3 ni(umsiúd) = #praenomen
67 4 d!e![kiti]ú![d] = #gentilicium
68 5 [mi](ínieís) = #patronymic
69 3;4;5 = @p1
70 6 . = #pc_word

```

Save Delete

Figure 4: ItAntDSL: textual data

as either words or names, uniquely identify them with an `@xml:id` attribute and, in the case of names, further specify them according to an appropriate taxonomy (*'praenomen'*, *'gentilicium'*, *'patronymic'*, etc.).

4 Application of ItAnt DSL to Faliscan

A linguistics graduate, proficient in epigraphy of the fragmentary languages of ancient Italy but with only basic skills in DH and particularly in text encoding, was selected to collaborate in the testing phase. Specifically, she was entrusted with five Faliscan inscriptions and tasked with encoding them both in

```

145- <tei:div type="edition" subtype="interpretative" xml:space="preserve">
146- <tei:div type="textpart" n="face_a" style="text-direction:r-to-l" rend="ductus:sinistrorse">
147- <tei:ab>
148- <tei:lb n="1" xml:lang="osc-Ital-x-oscetr" xml:id="Osc_3_1_1"/>
149- <tei:w xml:lang="osc-Ital-x-oscetr" xml:id="Osc_3_1_1_w_1">
150- <tei:expan><tei:abbr><tei:supplied reason="lost" evidence="previouseditor">m</tei:supplied></tei:abbr><tei:ex>edikúd</tei:ex></tei:expan>
151- </tei:w>
152- <tei:w xml:lang="osc-Ital-x-oscetr" xml:id="Osc_3_1_1_w_2">
153- <tei:expan><tei:abbr><tei:supplied reason="lost" evidence="previouseditor">t</tei:supplied></tei:abbr><tei:ex>úvtikúd</tei:ex></tei:expan>
154- </tei:w>
155- <tei:name type="praenomen" xml:lang="osc-Ital-x-oscetr" xml:id="Osc_3_1_1_w_3" ref="#p1">
156- <tei:expan><tei:abbr><tei:supplied reason="lost" evidence="previouseditor">ni</tei:supplied></tei:abbr><tei:ex>umsiú</tei:ex></tei:expan>
157- </tei:name>
158- <tei:name type="gentilicium" xml:lang="osc-Ital-x-oscetr" xml:id="Osc_3_1_1_w_4" ref="#p1">
159- <tei:unclear>de</tei:unclear>
160- <tei:supplied reason="lost" evidence="previouseditor">kiti</tei:supplied>
161- <tei:unclear>ú</tei:unclear><tei:supplied reason="lost" evidence="previouseditor">d</tei:supplied>
162- </tei:name>
163- <tei:name type="patronymic" xml:lang="osc-Ital-x-oscetr" xml:id="Osc_3_1_1_w_5" ref="#p1">
164- <tei:expan><tei:abbr><tei:supplied reason="lost" evidence="previouseditor">mi</tei:supplied></tei:abbr><tei:ex>iniéis</tei:ex></tei:expan>
165- </tei:name>
166- <tei:pc unit="word">.</tei:pc>
167- <!-- .... -->

```

Figure 5: XML-TEI/EpiDoc

XML-TEI/EpiDoc and through ItAntDSL. Although the case study lacks scientific relevance, it nonetheless provided interesting qualitative insights. The results can be summarized as follows: the time required for training was significantly shorter for learning the DSL compared to learning XML-TEI encoding; the time needed for encoding was markedly lower; documents produced via ItAntDSL are approximately three times more compact than EpiDoc documents.

LANGUAGE

#11: “Faliscan” (“Faliscan in Faliscan alphabet”)

Figure 6: Fragment of ItAntDSL metadata for a Faliscan inscription

During the text encoding phase, the need arose to introduce extensions to the original grammar, according to the planned workflow. The encoded documents (in Fig. 6 it is possible to see a couple of lines about the language of the inscriptions), processed by the ItAntDSL parser that generates an Abstract Syntax Tree (Fig. 7), produces XML files with a proprietary schema, as depicted in Fig. 8.

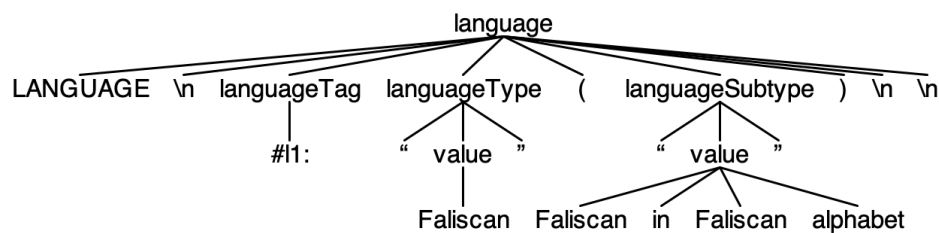


Figure 7: Example of Abstract Syntax Tree for ItAntDSL

To keep the language productions as compact as possible, shared information, which is typically repeated in the XML-TEI/EpiDoc files, is stored in a separated file. Fig. 9 shows a fragment of information related to the languages studied in this project, encoded in YAML and easily convertible to XML.

The XSLT transformation merges data from the XML file with proprietary schema with data from the YAML file converted to XML with a proprietary schema.

```

<language>
  LANGUAGE
  <languageTag>#1:</languageTag>
  <languageType>"<value>Faliscan</value>"</languageType>
  (<languageSubtype> "<value>Faliscan in Faliscan alphabet</value>"</languageSubtype>)
</language>

```

Figure 8: XML encoding with proprietary schema

<pre> list: type: "language" items: - key: "Faliscan" values: - type: "ident" content: "xfa" - type: "source" content: "https://iso639-3.sil.org/code/xfa" - key: "Faliscan in Faliscan alphabet" values: - type: "ident" content: "xfa-Ital-x-xfafal" - type: "source" content: "https://www.prin-italia-antica.unifi.it/" - type: "ana" content: "https://unicode.org/iso15924/iso15924-codes.html" </pre>		<pre> <list xmlns="http://itant.eu" type="language"> <item key="Faliscan"> <value type="ident">xfa</value> <value type="source">https://iso639-3.sil.org/code/xfa</value> </item> <item key="Faliscan in Faliscan alphabet"> <value type="ident">xfa-Ital-x-xfafal</value> <value type="source">https://www.prin-italia-antica.unifi.it/</value> <value type="ana">https://unicode.org/iso15924/iso15924-codes.html</value> </item> </list> </pre>
--	---	--

Figure 9: Fragment of the YAML file with look-up information

```

<xsl:template match="//dsl:start/dsl:language/dsl:languageType/dsl:value|/dsl:start/dsl:language/dsl:languageSubtype/dsl:value">
  <tei:language>
    <xsl:variable name="key" select="."/ >
    <xsl:variable name="languageIdent" select="document('database.xml')/dsl:data/dsl:list/dsl:item[@key=$key]/dsl:value[@type='ident']"/ >
    <xsl:variable name="languageSource" select="document('database.xml')/dsl:data/dsl:list/dsl:item[@key=$key]/dsl:value[@type='source']"/ >
    <xsl:variable name="languageAna" select="document('database.xml')/dsl:data/dsl:list/dsl:item[@key=$key]/dsl:value[@type='ana']"/ >
    <xsl:attribute name="ident">
      <xsl:value-of select="$languageIdent"/ >
    </xsl:attribute>
    <xsl:if test="$languageSource">
      <xsl:attribute name="source">
        <xsl:value-of select="$languageSource"/ >
      </xsl:attribute>
    </xsl:if>
    <xsl:if test="$languageAna">
      <xsl:attribute name="ana">
        <xsl:value-of select="$languageAna"/ >
      </xsl:attribute>
    </xsl:if>
    <xsl:value-of select="$key"/ >
  </tei:language>
</xsl:template>

```

Figure 10: Fragment of the XSLT file

The resulting XML-TEI/EpiDoc fragment is visible in Fig. 11.

```

<tei:langUsage>
  <tei:language ident="xfa" source="https://iso639-3.sil.org/code/xfa">Faliscan</tei:language>
  <tei:language ident="xfa-Ital-x-xfafal"
    source="https://www.prin-italia-antica.unifi.it/"
    ana="https://unicode.org/iso15924/iso15924-codes.html">Faliscan in Faliscan alphabet</tei:language>
</tei:langUsage>

```

Figure 11: XML-TEI/EpiDoc output

5 Conclusions

In this paper, we have presented a novel approach to addressing the challenges associated with encoding fragmentary languages of ancient Italy in digital formats, in ways that are user-friendly and meaningful within the discipline. Through the development and application of a Domain-Specific Language (DSL) called ItAntDSL, tailored to the needs of historical linguists and epigraphists, we have improved in efficiency, compactness, and ease of use compared to traditional XML-TEI/EpiDoc encoding methods.

Our experiment with encoding five Faliscan inscriptions using both ItAntDSL and XML-TEI/EpiDoc seems to reveal that the training time required for mastering the DSL is significantly shorter, while the encoding process itself is markedly more efficient. Furthermore, documents produced via ItAntDSL were approximately three times more compact than their EpiDoc counterparts, making them much more readable and accessible for humans as well as significantly reducing file size and complexity. The successful application of ItAntDSL in encoding ancient inscriptions underscores its potential as a powerful tool for digital epigraphy and historical linguistics. It advocates for an integration within platforms such as DigItAnt, thereby permitting a full editing experience from within a single online environment.

While the ItAnt project provided a valuable opportunity to develop methods and tools to facilitate the encoding activities of epigraphists, CLARIN offers not only the infrastructure to deposit research data, but also the means to disseminate and share new practices adequate to the domain of epigraphic studies. In recent years, collaborative efforts within the CLARIN Knowledge Centre for Digital and Public Textual Scholarship (DiPText-KC)¹⁸ have resulted in significant contributions to digital humanities projects. These efforts, involving the Venice Center for Digital and Public Humanities (VeDPH), the Institute for Computational Linguistics (CNR-ILC), and CLARIN-IT center ILC4CLARIN have addressed various kinds of resources, including DH projects related to collections of literary texts (Boschetti et al., 2021), and collections of epigraphic sources (Vagionakis et al., 2022).

5.1 Future works

The know-how acquired in the process of encoding and annotating inscriptions through ItAntDSL will be shared within our target disciplinary scientific community through the CLARIN Knowledge Center DiPText. Knowledge sharing may take the form of video tutorials, webinars, and/or workshops, and will thus continue also after the end of the ItAnt project.

The corpora containing the digital editions of the inscriptions will be deposited into CLARIN as soon as they are finalized. Additionally, the ItAnt DSL has already been deposited (Boschetti & Rigobianco, 2023) and it will be updated to integrate any improvements.

Regarding future improvements to the DigItAnt platform, we will explore the possibility of integrating the ItAntDSL system into the platform's interface, to allow users to create or revise editions of the inscriptions via the DSL directly within the DigItAnt web environment.

Acknowledgments

This work is supported by the Italian Ministry of the University and Research with the Italian National Strategic Research Grant PRIN 2017XJLE8J for the project: *Languages and Cultures of Ancient Italy. Historical Linguistics and Digital Models*. The project involves a consortium comprising the Ca' Foscari University of Venice, the University of Florence, and the Institute for Computational Linguistics "A. Zampolli" of the National Research Council of Italy. It also benefits from collaboration with the Venice Center for Digital and Public Humanities (VeDPH), which is part of the Department of Humanities of Ca' Foscari University of Venice (<https://www.unive.it/pag/39287>).

The project is also supported by and contributing to the CLARIN-IT research infrastructure for the Humanities and Social Sciences.

¹⁸<https://diptext-kc.clarin-it.it/>

References

- Bellandi, A. (2019). LexO-server: REST services for linguistic linked data in OntoLex-lemon [ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa]. <http://hdl.handle.net/20.500.11752/ILC-1004>
- Boschetti, F., Del Grosso, A. M., & Spinazzè, L. (2021). La galassia musisque deoque: Storia e prospettive. In *Paulo maiora canamus - raccolta di studi per Paolo Mastandrea* (pp. 405–419, Vol. 32). Edizioni CaFoscari. <https://edizionicafoscari.unive.it/media/pdf/books/978-88-6969-558-2/978-88-6969-558-2-ch-26.pdf>
- Boschetti, F., & Rigobianco, L. (2023). ItAntDSL [ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa]. <http://hdl.handle.net/20.500.11752/ILC-1003>
- Krummrey, H., & Panciera, S. (1980). Criteri di edizioni e segni diacritici. In G. Bevilacqua (Ed.), *Tituli* (Vol. 2). Edizioni di Storia e Letteratura.
- Mallia, M., Bellandi, A., Tommasi, A., Zavattari, C., Bandini, M., & Quochi, V. (2023). EpiLexO [ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa]. <http://hdl.handle.net/20.500.11752/ILC-1005>
- Marinetti, A. (2020). Venetico. *Palaeohispanica. Revista sobre lenguas y culturas de la Hispania Antigua*, (20), 367–401. <https://doi.org/10.36707/palaeohispanica.v0i20.374>
- Murano, F., Quochi, V., Del Grosso, A. M., Rigobianco, L., & Zinzi, M. (2023). Describing Inscriptions of Ancient Italy. The ItAnt Project and Its Information Encoding Process. *Journal on Computing and Cultural Heritage*, 16, 1–14. <https://doi.org/10.1145/3593431>
- Neill, I., & Schmidt, D. (2021). SPEEDy. A Practical Editor for Texts Annotated with Standoff Properties. *Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing*, 15, 45.
- Parr, T. (2009). *Language implementation patterns: create your own domain-specific and general programming languages*. The Pragmatic Bookshelf.
- Parr, T. (2013). *The definitive ANTLR 4 reference*. The Pragmatic Bookshelf.
- Pocchetti, P. (2017). The documentation of Italic. In J. Klein, B. Joseph, & M. Fritz (Eds.), *Handbook of Comparative and Historical Indo-European Linguistics* (pp. 733–742, Vol. 2). De Gruyter Mouton. <https://www.degruyter.com/document/doi/10.1515/9783110523874-001/html>
- Prag, J. R. W., & Chartrand, J. (2019). I. Sicily: Building a Digital Corpus of the Inscriptions of Ancient Sicily. In A. D. Santis & I. Rossi (Eds.), *Crossing Experiences in Digital Epigraphy: From Practice to Discipline* (pp. 240–252). De Gruyter Open Poland. <https://doi.org/10.1515/9783110607208-020>
- Quochi, V., Bellandi, A., Khan, F., Mallia, M., Murano, F., Piccini, S., Rigobianco, L., Tommasi, A., & Zavattari, C. (2022). From Inscriptions to Lexicon and Back: A Platform for Editing and Linking the Languages of Ancient Italy. *Proceedings of Second Workshop on Language Technologies for Historical and Ancient Languages LT4HALA 2022*, 59–67.
- Quochi, V., Bellandi, A., Mallia, M., Tommasi, A., & Zavattari, C. (2022). Supporting Ancient Historical Linguistics and Cultural Studies with EpiLexO. *CLARIN Annual Conference Proceedings*, 39.
- Sarullo, G. (2016). The encoding challenge of the ILA project. In A. E. Felle & A. Rocco (Eds.), *Off the beaten track. epigraphy at the borders* (pp. 15–17). Archaeopress. <https://www.archaeopress.com/Archaeopress/download/9781784913229.pdf#page=25>
- Spadini, E., & Turska, M. (2019). XML-TEI Stand-off Markup: One Step Beyond. *Digital Philology: A Journal of Medieval Cultures*, 8(2), 225–239.
- Vagionakis, I., Del Gratta, R., Boschetti, F., Baroni, P., Del Grosso, A. M., Mancinelli, T., & Monachini, M. (2022). ‘Cretan Institutional Inscriptions’ Meets CLARIN-IT. *CLARIN Annual Conference*, 139–150.
- Zenzaro, S., Grosso, A. M. D., Boschetti, F., & Ranocchia, G. (2022). Verso la definizione di criteri per valutare soluzioni di scholarly editing digitale: Il caso d’uso GreekSchools. In F. Ciracì, G.

Miglietta, & C. Gatto (Eds.), *Aiucd 2022 proceedings* (pp. 20–25). <https://amsacta.unibo.it/id/eprint/6848/>