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 activity 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, by increasing the human readability of the encoded data without sacrificing the compliance to standard models and formats. Such a work is framed in the context of an Italian National collaborative research project devoted to the study of the languages and cultures of ancient Italy, witnessed by a digital collection of inscriptions. The platform developed within this project offers an interesting use case and motivation for experimenting with DSLs for the creation of the needed digital critical editions. After explaining the DSL grammar definition process, we finally, test the applicability of the DSL grammar to 5 example inscriptions in the Faliscan language.

1 Introduction

The recovery, digitisation, and sharing of knowledge relating to ancient fragmentary languages and their cultures is a primary objective, which, at the same time, represents a challenge for both historical linguistics and digital humanities.

Fragmentary languages are dead languages attested through a highly restricted corpus of 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 this restriction, the knowledge which may be obtained is 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 a partial, uncertain and quantitatively scare knowledge prevents the use of AI or machine learning techniques, as currently designed, and requires adapting available language technology tools, which may only be achieved through the cooperation between historical linguists and language technologists.

A very first, fundamental and by all means non-trivial stage in this direction is the creation of solid machine-actionable digital scholarly editions of the inscriptions and their linguistic content. Recently, the ILA project (Sarullo, 2016) has taken the first steps in the direction of adapting the TEI-EpiDoc standard to an epigraphically attested fragmentary language such as archaic Latin (7th-5th century BC). Furthermore, among others, the i.Sicily digital corpus (Prag & Chartrand, 2019) shall be mentioned, which 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 adequate digital treatment of these languages must be taken up in order to preserve their documentation and knowledge and make them widely accessible. However, digitising scholarly editions proves to be a a time-consuming and unfriendly task for many scholars. This contribution introduces and discusses on a method and technological solution based on Domain Specific Languages (DSLs hereafter), to facilitate such a task.

The paper is organised as follows: section 2 describes the project that motivated this work, and the online platform which will consume the produced critical editions. Here the connection to the CLARIN infrastructure is also made explicit. Section 3 gives a quick introduction to DSLs and their advantage in Digital Humanities (DH) contexts, before presenting in detail the specific grammar designed for the languages of ancient Italy. Section 4 instead demonstrates the applicability of the DSL grammar to 5 example inscriptions in the Faliscan language. Section 5, finally, wraps up and indicates some 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 an initiative funded by the Italian Ministry of University and Research and 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[1]. This project aims at investigating the languages of Ancient Italy combining the methods of historical linguistics with digital technologies specifically designed to create a set of interrelated resources, particularly critical digital editions of inscriptions, lexica and bibliographies.

With the sole exception of Roman Latin, the languages of Ancient Italy (8th century BC-1st cen- tury 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 the digital tools, practices, and methodologies of digital epigraphy and computational lexicography to the highly fragmentary nature of such a documentation. In particular, 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 difference in their documentation and their belonging to linguistic (sub)groups which are diverse as regards their genetic classification (Poccetti, 2017).

The main objectives of the project are to create and interlink one another a digital archive of (critical editions of) inscriptions, a multilingual computational lexicon, and a bibliographic dataset of all relevant cited works in FRBRoo/LRMoo[2]. With regard to the digital archive, the inscriptions are being encoded in XML according the XML-TEI/EpiDoc schema[3]. Furthermore, the edition of the inscriptions is enriched with standard metadata, thus allowing for an accurate description of each of them as both a linguistic and a material objects.

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: i.e. LOD-compliant lexica natively interlinked with a family of related resources: critical editions of inscriptions, citations and bibliographic references, plus other external available salient vocabularies and lexicons.

Assuming that more intuitive disciplinary editing tools can simplify the work of philologists and historical linguists in the management of lexical and linguistic knowledge about ancient languages, the DigItAnt platform is meant to assist scholars with the encoding lexical information of ancient languages and in linking it to other relevant (re-)sources according to the semantic web principles. Particularly central to the platform is the linking of lexical and morphological forms to their attestations in the texts encoded in TEI-EpiDoc digital scholarly editions of relevant inscriptions. Thus, lexicon creation is at the heart of the editing platform, which enables scholars to enrich them with actionable links to related inscription, to cited bibliographic items, and to other external datasets[4].

Digital critical editions of inscriptions, albeit crucial for the scholar to consult, currently have an ancillary role in the editing platform as they are considered instrumental to lexicon encoding. Inscriptions are presently assumed to be encoded independently of the platform, in XML according to the TEI/EpiDoc

[1] https://www.prin-italia-antica.unifi.it/

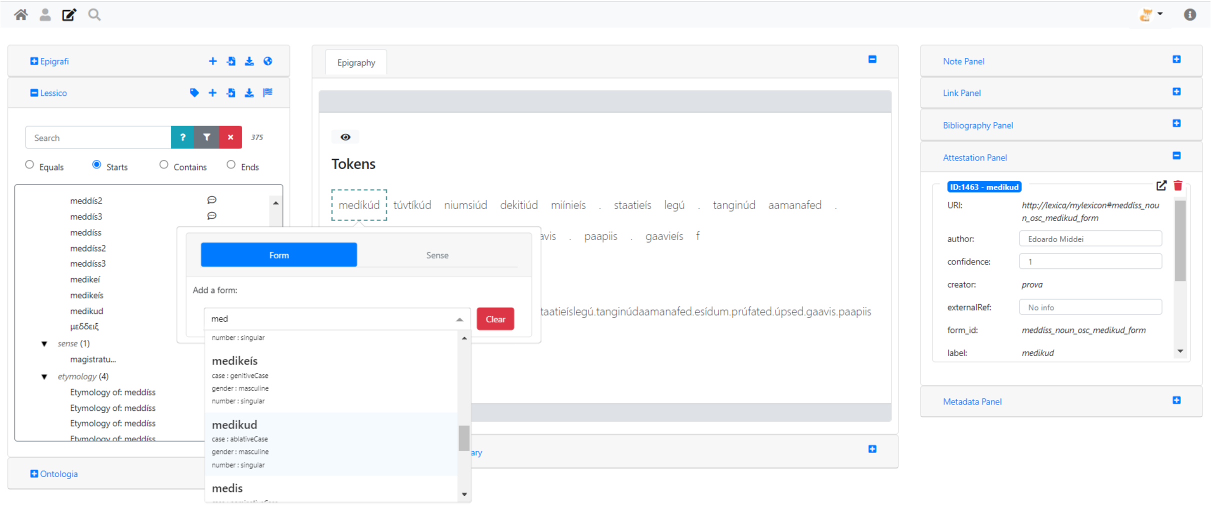
[2] https://cidoc-crm.org/frbroo/

[3] http: //www.stoa.org/epidoc/gl/latest/

[4] See Quochi, Bellandi, Mallia, et al. (2022) and Quochi, Bellandi, Khan, et al. (2022) for details on the platform).

format, the de-facto standard for digital epigraphic projects. Digital editions of inscriptions are thus considered as external datasets that the platform can ingest.

Within ItAnt, we have thus experimented with the use of a DSL, as an alternative tool to the more commonly used Oxygen XML editor, which might offer scholars a lighter and more intuitive way to produce their digital editions. Thus, the editions encoded with ItAntDSL, and then converted to EpiDoc-XML as described below, can be subsequently ingested by the platform for linking and exploration purposes. In details, the basic workflow would be such that the (historical linguist) user uploads one or more EpiDoc XML documents into the platform so that (s)he can link the exact text loci to either existing or newly created lexical items (as exemplified in Fig. 1, see also Quochi, Bellandi, Khan, et al. (2022)).



URI: http://lexica/mylexicon#meddíss_noun_osc_medikud_form
author: Edoardo Middei
confidence: 1
creator: prova
externalRef: No info
form_id: meddíss_noun_osc_medikud_form
label: medikud

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 conventions is also possible, both in the editor (Epilexo) and in the exploration platform (DigItAnt-search), as shown in Fig. 2[5].

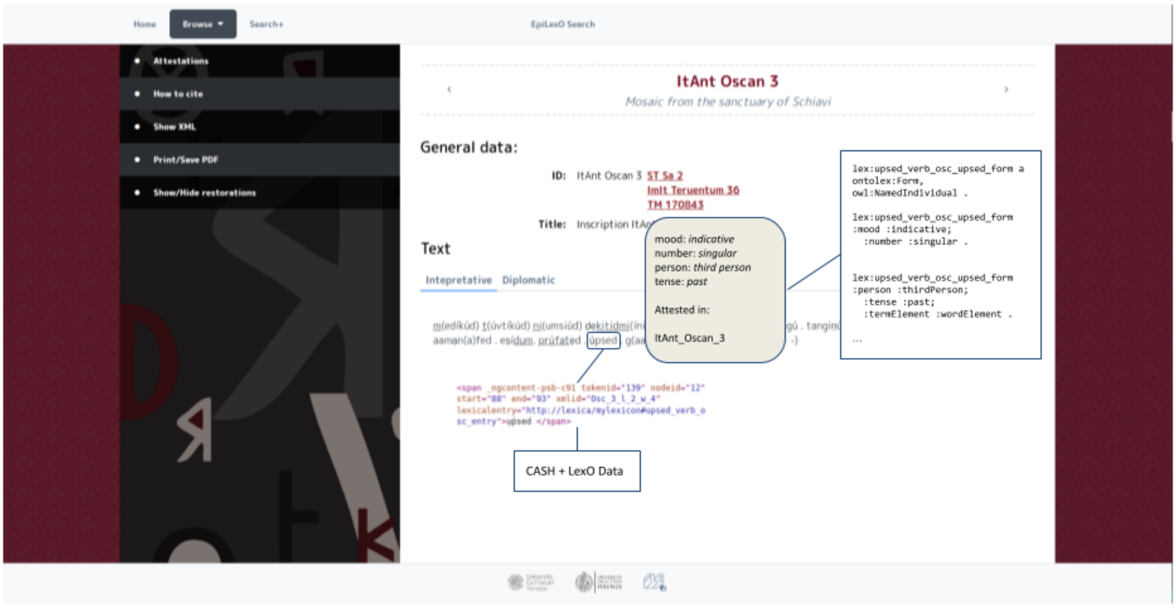
2.2 The Digital Edition of the Inscriptions

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 ancient documents according to the Leiden Conventions. In particular, XML-TEI/EpiDoc provides mark-ups 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).

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 in adding tags to the standard 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 inscriptions [6]. 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.

[5] The 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.

[6] On the peculiar Venetic syllabic punctuation see Marinetti (2020)



ItAnt Oscan 3

Mosaic from the sanctuary of Schiavi

General data:

ID: ItAnt Oscan 3 ST Sa 2
Imlt Teruentum 36
TM 170843

Title: Inscription ItAn

Text

m(edíkúd) t(úvtíkúd) ni(umsiúd) dekitidmi(ín … gú . tangin(
aaman(a)fed . esídum. prúfated . úpsed . g(aa … -)

```
<span _ngcontent-psb-c91 tokenid="139" nodeid="12"
start="88" end="93" xmlid="Osc_3_l_2_w_4"
lexicalentry="http://lexica/mylexicon#upsed_verb_o
sc_entry">upsed </span>
```

mood: *indicative*
number: *singular*
person: *third person*
tense: *past*

Attested in:

ItAnt_Oscan_3

```
lex:upsed_verb_osc_upsed_form a
ontolex:Form,
owl:NamedIndividual .

lex:upsed_verb_osc_upsed_form
:mood :indicative;
  :number :singular .


lex:upsed_verb_osc_upsed_form
:person :thirdPerson;
  :tense :past;
  :termElement :wordElement .

...
```

Figure 2: Text-lexicon data mesh-up in the Exploration platform

Additionally, a major problem with the 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 provision of lexical information is dealt with in the companion lexicon. 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.

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 *The Getty Research Institute for the object type*[7], material, and writing technique [8], the EAGLE vocabulary for the type of inscriptions (dedicatory, funerary, etc.)[9], and Pleiades and GeoNames for ancient and modern names respectively[10]. In addition, Trismegistos IDs are used, when available, to identify the text[11] and bibliographical records are also linked through a specific library built up by using Zotero[12].

[7] 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.

[8] https://www.getty.edu/research/tools/vocabularies/aat/

[9] https://www.eagle-network.eu/resources/vocabularies/typeins/

[10] https://pleiades.stoa.org/; https://www.geonames.org/

[11] https://www.trismegistos.org/tm/

[12] https://www.zotero.org/groups/2552746/

2.3 Relation with CLARIN

Part of the mission of the ItAnt project is to contribute and integrate as far as possible data and tools into European Research Infrastructures for the Humanities and social sciences, particularly to CLARIN. As such, since it started, ItAnt has been a project of interest for CLARIN-IT, also for its potential contribution to the involvement of the community of language historians. At the end of the project funding, the hosting of the platform will be moved from the project internal cloud space to the ILC4CLARIN center, which will offer it as a sustainable open service. Not only the platform but all software component will be preserved and distributed through CLARIN, for others to re-use. The ILC4CLARIN repository[13] already stores a copy of LexO-server[14], EpiLexO[15] and ItAntDSL[16]. At the end of the project all software, inscription corpora and ancient lexicons will be preserved, discoverable and consumable via CLARIN channels.

Furthermore, because of the ItAnt focus on publishing also LLOD compatible versions of the data, it will contribute to the development of a CLARIN(-IT) LLOD Hub. In this respect, DigItAnt is a candidate use case for one of the pilot projects to be developed in the context of an ongoing big Italian infrastructural endeavour, the Humanities and Heritage Italian Open Science Cloud (H2IOSC)[17].

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 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 forth one, whereas a named entity defined by *praenomen* (partially conjectured), *gentilicium* and *patronymicus* may extend from the forth 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,

[13] https://ilc4clarin.ilc.cnr.it/

[14] http://hdl.handle.net/20.500.11752/ILC-1004

[15] http://hdl.handle.net/20.500.11752/ILC-1005

[16] http://hdl.handle.net/20.500.11752/ILC-1003

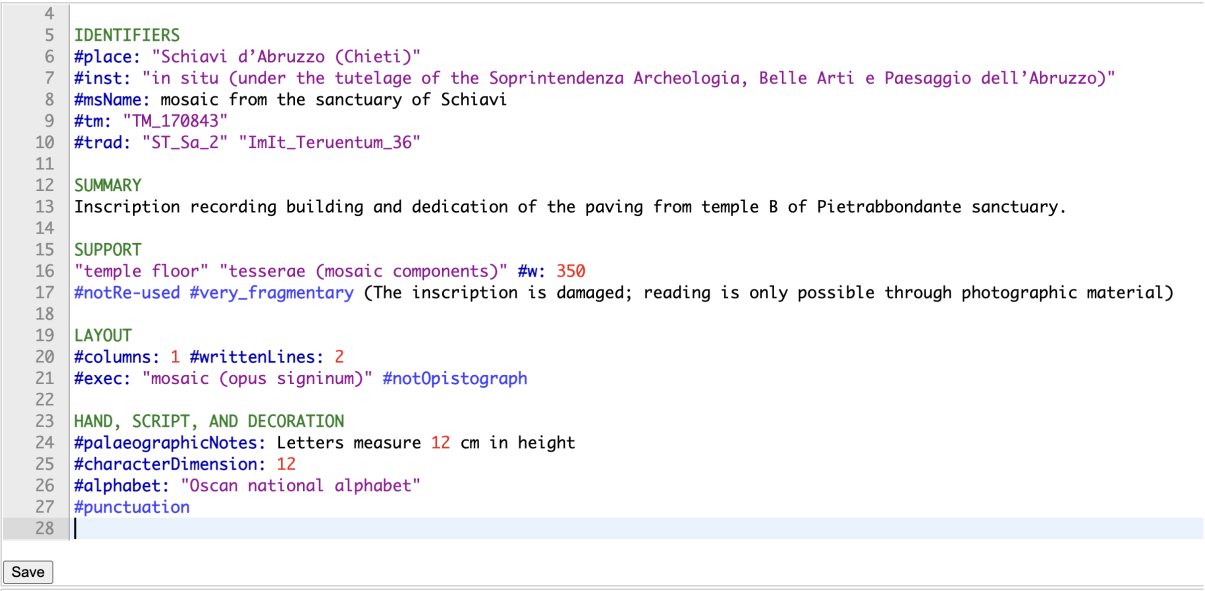
[17] https://www.h2iosc.cnr.it/home/

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 to 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 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 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 github[18]. 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.

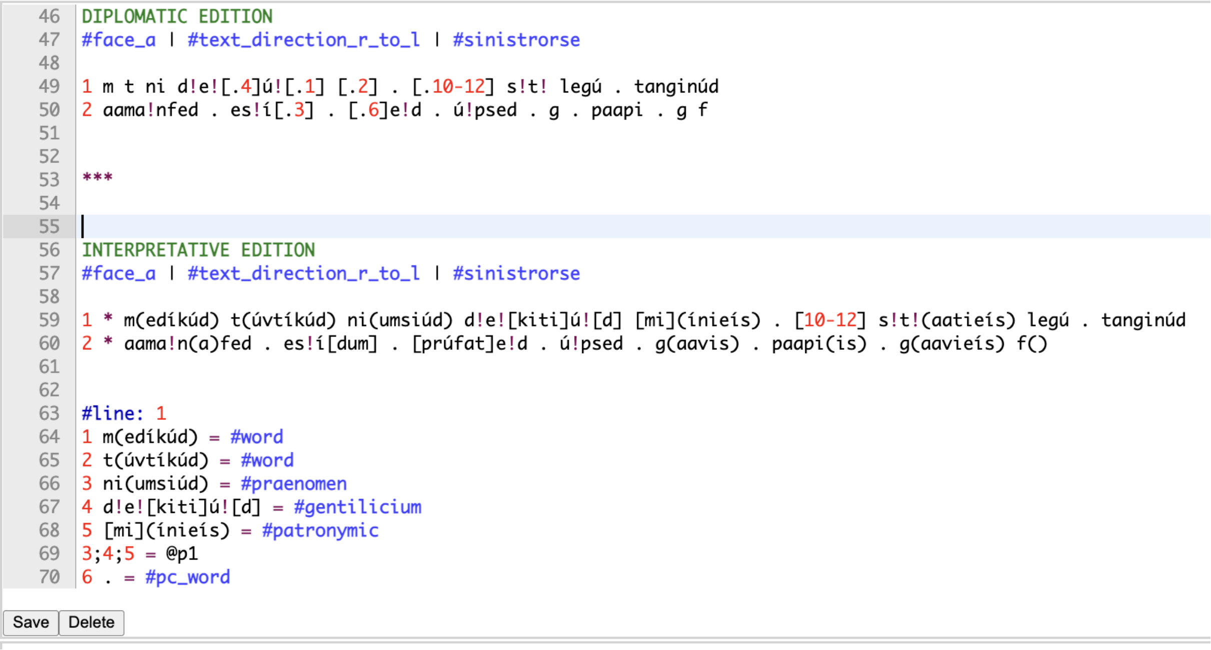


```

IDENTIFIERS
#place: "Schiavi d'Abruzzo (Chieti)"
#inst: "in situ (under the tutelage of the Soprintendenza Archeologia, Belle Arti e Paesaggio dell'Abruzzo)"
#msName: mosaic from the sanctuary of Schiavi
#tm: "TM_170843"
#trad: "ST_Sa_2" "ImIt_Teruentum_36"

SUMMARY
Inscription recording building and dedication of the paving from temple B of Pietrabbondante sanctuary.

SUPPORT
"temple floor" "tesserae (mosaic components)" #w: 350
#notRe-used #very_fragmentary (The inscription is damaged; reading is only possible through photographic material)

LAYOUT
#columns: 1 #writtenLines: 2
#exec: "mosaic (opus signinum)" #notOpistograph

HAND, SCRIPT, AND DECORATION
#palaeographicNotes: Letters measure 12 cm in height
#characterDimension: 12
#alphabet: "Oscan national alphabet"
#punctuation

```

Figure 3: ItAntDSL: metadata

Then, a chain of xquery scripts and XSLT stylesheets transforms XML-Itant documents into XML-

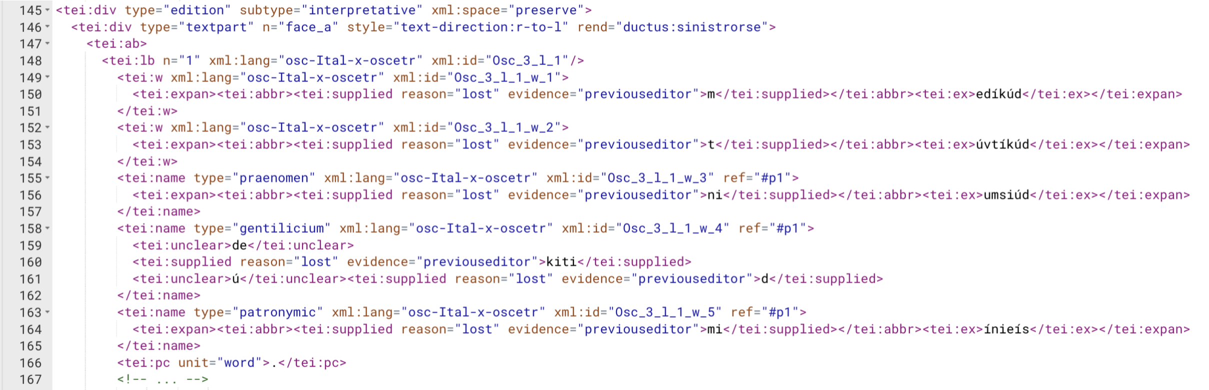
[18] https://github.com/CoPhi/itantdsl/



```
DIPLOMATIC EDITION
#face_a | #text_direction_r_to_l | #sinistrorse

1 m t ni d!e![.4]ú![.1] [.2] . [.10-12] s!t! legú . tanginúd
2 aama!nfed . es!í[.3] . [.6]e!d . ú!psed . g . paapi . g f


***


INTERPRETATIVE EDITION
#face_a | #text_direction_r_to_l | #sinistrorse

1 * m(edíkúd) t(úvtíkúd) ni(umsiúd) d!e![kiti]ú![d] [mi](ínieís) . [10-12] s!t!(aatieís) legú . tanginúd
2 * aama!n(a)fed . es!í[dum] . [prúfat]e!d . ú!psed . g(aavis) . paapi(is) . g(aavieís) f()


#line: 1
1 m(edíkúd) = #word
2 t(úvtíkúd) = #word
3 ni(umsiúd) = #praenomen
4 d!e![kiti]ú![d] = #gentilicium
5 [mi](ínieís) = #patronymic
3;4;5 = @p1
6 . = #pc_word
```

Figure 4: ItAntDSL: textual data

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-db. A sample of the final result is shown if Fig. 5.

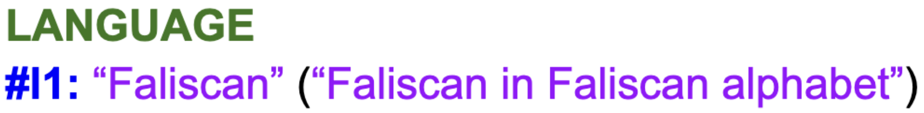


```
<tei:div type="edition" subtype="interpretative" xml:space="preserve">
  <tei:div type="textpart" n="face_a" style="text-direction:r-to-l" rend="ductus:sinistrorse">
    <tei:ab>
      <tei:lb n="1" xml:lang="osc-Ital-x-oscetr" xml:id="Osc_3_l_1"/>
        <tei:w xml:lang="osc-Ital-x-oscetr" xml:id="Osc_3_l_1_w_1">
          <tei:expan><tei:abbr><tei:supplied reason="lost" evidence="previouseditor">m</tei:supplied></tei:abbr><tei:ex>edíkúd</tei:ex></tei:expan>
        </tei:w>
        <tei:w xml:lang="osc-Ital-x-oscetr" xml:id="Osc_3_l_1_w_2">
          <tei:expan><tei:abbr><tei:supplied reason="lost" evidence="previouseditor">t</tei:supplied></tei:abbr><tei:ex>úvtíkúd</tei:ex></tei:expan>
        </tei:w>
        <tei:name type="praenomen" xml:lang="osc-Ital-x-oscetr" xml:id="Osc_3_l_1_w_3" ref="#p1">
          <tei:expan><tei:abbr><tei:supplied reason="lost" evidence="previouseditor">ni</tei:supplied></tei:abbr><tei:ex>umsiúd</tei:ex></tei:expan>
        </tei:name>
        <tei:name type="gentilicium" xml:lang="osc-Ital-x-oscetr" xml:id="Osc_3_l_1_w_4" ref="#p1">
          <tei:unclear>de</tei:unclear>
          <tei:supplied reason="lost" evidence="previouseditor">kiti</tei:supplied>
          <tei:unclear>ú</tei:unclear><tei:supplied reason="lost" evidence="previouseditor">d</tei:supplied>
        </tei:name>
        <tei:name type="patronymic" xml:lang="osc-Ital-x-oscetr" xml:id="Osc_3_l_1_w_5" ref="#p1">
          <tei:expan><tei:abbr><tei:supplied reason="lost" evidence="previouseditor">mi</tei:supplied></tei:abbr><tei:ex>ínieís</tei:ex></tei:expan>
        </tei:name>
        <tei:pc unit="word">.</tei:pc>
        <!-- ... -->
```

Figure 5: XML-TEI/EpiDoc

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 XML-TEI 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
#l1: “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 is possible to see a couple of lines about the language of the inscriptions), processed by ItAntDSL parser, which generates an Abstract Syntax Tree (Fig. 7), produces XML files with a proprietary schema, as depicted in Fig. 8.

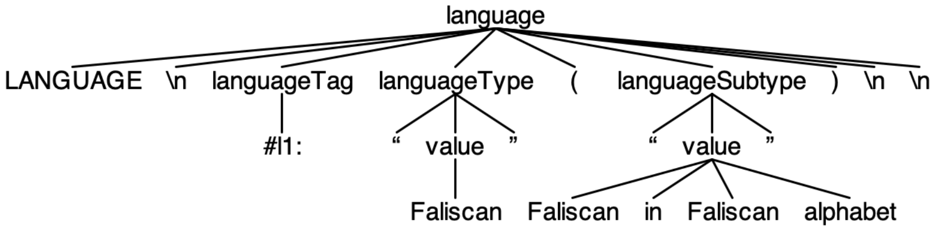
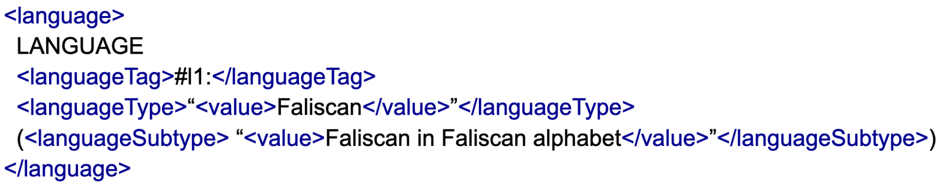


Figure 7: Example of Abstract Syntax Tree for ItAntDSL

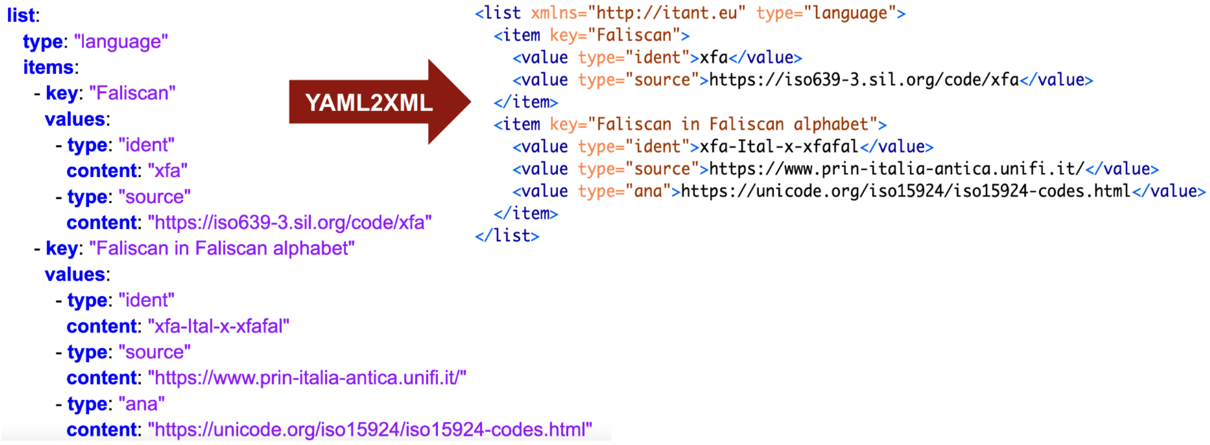


```
<language>
  LANGUAGE
  <languageTag>#l1:</languageTag>
  <languageType>“<value>Faliscan</value>”</languageType>
  (<languageSubtype> “<value>Faliscan in Faliscan alphabet</value>”</languageSubtype>)
</language>
```

Figure 8: XML encoding with proprietary scheme

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

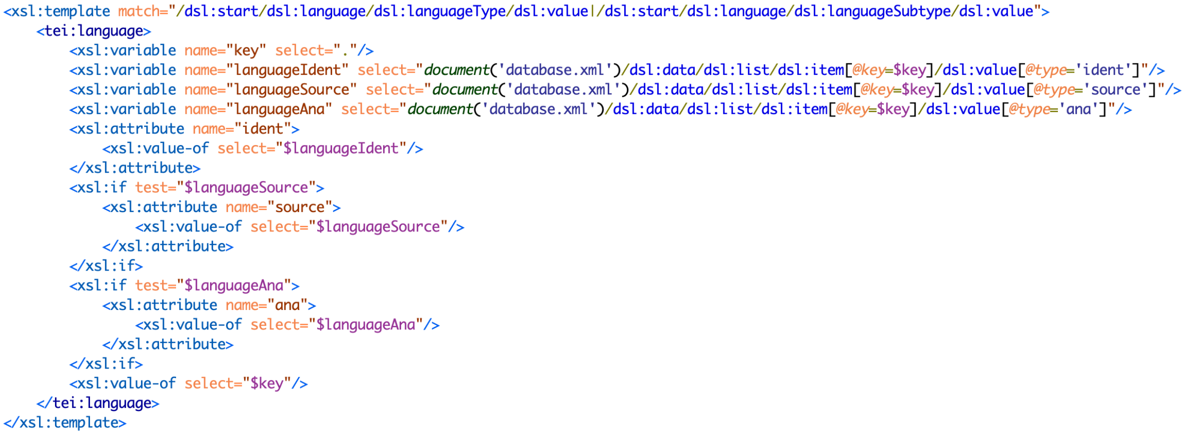
The XSLT transformation merges data from the XML file with proprietary scheme with data from the YAML file converted in XML with a proprietary scheme.



```
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"
```

```
<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>
```

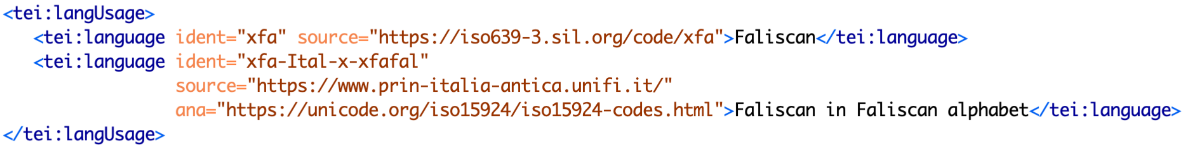
Figure 9: Fragment of the YAML file with look-up information



```
<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

VeDPH, CNR-ILC and ILC4CLARIN in the last years are collaborating on DH projects related to many kinds of resources, such as collections of literary texts (Boschetti et al., 2021) and collections of epigraphic sources (Vagionakis et al., 2022). ItAnt provides a good opportunity to develop methods and tools to facilitate the encoding activities of the epigraphists, which must deal with complex entangled data. CLARIN provides not only the infrastructure to deposit the research data, but also the instruments to share new practices adequate to the domain of the epigraphic studies.

5.1 Future works

The know-how related to the annotation of inscriptions through ItAntDSL will be shared through the Digital and Public Textual Scholarship Knowledge Centre[19] (DiPText-KC) of CLARIN. Videotutorials and other initiatives, such as webinars and workshops, are planned towards the end of the project and after.

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 is also supported by and contributing to the CLARIN-IT research infrastructure for the Humanities and Social Sciences.
All authors have read and agreed to the submitted extended version of the manuscript.

References

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

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.

Poccetti, 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.

[19] https://diptext-kc.clarin-it.it/

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/