

Domain-Specific Languages for Epigraphy

the Case of ItAnt

Federico Boschetti, Luca Rigobianco, Valeria Quochi

CNR-ILC

CLARIN Conference 2023

Leuven, 17 October 2023

Introduction to ItAnt

The Intersection of Epigraphy and Digital Technologies

- In the last decades, digital epigraphy has grown and several projects of digitisation of inscriptions have been launched.
- Many of these projects make use of the TEI-EpiDoc XML encoding standard for ensuring machine readability and interoperability.

Project Overview

Languages and Cultures of Ancient Italy. Historical Linguistics and Digital Models

- 20/1/2020 - 20/7/2024
- ItAnt is a Project of Relevant National Interest (PRIN) for the Italian Ministry of University and Research
- The project involves a consortium composed by the University Ca' Foscari of Venice, the University of Florence, and the Institute for Computational Linguistics "A. Zampolli" of the National Research Council of Italy

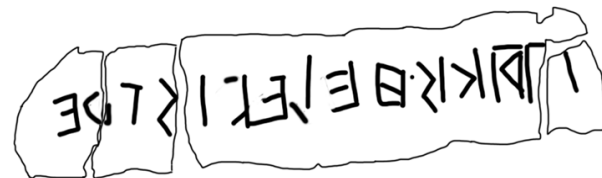
Project Overview

Languages and Cultures of Ancient Italy. Historical Linguistics and Digital Models

- The project aims at investigating the languages of Ancient Italy, combining the classical methods of historical linguistics with the setting up of digital and computational technologies specifically designed to create a set of interrelated and interoperable digital resources.
- The main objectives of the project are to create and interlink a digital archive of critically edited inscriptions, a multilingual computational lexicon, relevant bibliography and citations.

Challenges

Adapting Digital Tools for Fragmentary Languages



- With the sole exception of Roman Latin, the languages of Ancient Italy (8th century BC-1st century AD) are fragmentary languages, that is to say dead languages attested through a highly restricted corpus of texts.
- Their evidence consists almost exclusively of epigraphic texts, which often present problems relating to the reading, segmentation into words, linguistic analysis, and interpretation.

Digital Edition of Inscriptions

XML-TEI/EpiDoc Schema and Customisations

- The customisation has essentially consisted of adding tags to the standard set in order to describe features occurring in the inscriptions which may be considered relevant to the study of the languages of ancient Italy, such as:
 - a `<tei:rs>` element with a `@type="wordDivision"` within the `<scriptNote>` element for encoding the use of scriptio continua or punctuation, blank spaces, and mixed systems for word division;
 - a `<tei:rs>` element with a `@type="syllabicPunctuation"` within the `<scriptNote>` element for encoding the Venetic syllabic punctuation;
 - etc.

Data Integration and Semantic Web readiness

- TEI-EpiDoc documents make use of several common vocabularies for metadata and geographical information (e.g. GeoNames, Pleiades, Art & Architecture Thesaurus, iDAI. Thesauri, EAGLE vocabularies), and need to be explicitly linked to lexica and bibliographic datasets;
- Lexical encoding conforms to Ontolex-lemon and extensions;
- LexInfo is used for linguistic descriptors;
- Lexical data is linked to external LOD resources, e.g. the LiLa Knowledge Base, bibliographic data, texts spans in the ItAnt scholarly editions.

The Complexity of XML-TEI Encoding

- TEI-EpiDoc is a *de facto* standard in the subdomain of Digital Epigraphy
- It ensures machine actionability and interoperability
- The drawback is that an EpiDoc document is verbose and difficult to be read by a traditional epigraphist

Introducing ItAntDSL

- ItAntDSL is a Domain-Specific Language for Epigraphy focused on the needs of the ItAnt community, which is mainly constituted by traditional epigraphists
- As a DSL, the language is defined by a Context-Free Grammar, created by computational linguists according to the requirements of the epigraphists

```
5 IDENTIFIERS
6 #place: "Schiavi d'Abruzzo (Chieti)"
7 #inst: "in situ (under the tutelage of the Soprintendenza Archeologia, Belle Arti e Paesaggio dell'Abruzzo)"
8 #msName: mosaic from the sanctuary of Schiavi
9 #tm: "TM_170843"
10 #trad: "ST_Sa_2" "ImIt_Teruentum_36"
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 #paleographicNotes: Letters measure 12 cm in height
25 #characterDimension: 12
26 #alphabet: "Oscan national alphabet"
27 #punctuation
28
```

Save Delete

```
46 DIPLOMATIC EDITION
47 #face_a | #text_direction_r-to_l | #sinistrorse
48
49 1 m t ni die[.4]ú[.1] [.2] . [10-12] sit! legú . tanginú
50 2 aama!nfed . es!i[.3] . [.6]eid . úpsed . g . paapt . g f
51
52 ***
53
54
55
56 INTERPRETATIVE EDITION
57 #face_a | #text_direction_r-to_l | #sinistrorse
58
59 1 * m(edikúd) t(úvtikúd) ni(umsiúd) die! [kiti]ú! [d] [mi](inieis) . [10-12] sit!(aatieis) legú . tanginú
60 2 * aama!n(a)fed . es!i[dum] . [prúfat]eid . úpsed . g(aavis) . paapt(is) . g(aavieis) f()
61
62
63 #line: 1
64 1 m(edikúd) = #word
65 2 t(úvtikúd) = #word
66 3 ni(umsiúd) = #praenomen
67 4 die! [kiti]ú! [d] = #gentilicium
68 5 [mi](inieis) = #patronymic
69 3;4;5 = @p1
70 6 . = #pc_word
```

Save Delete

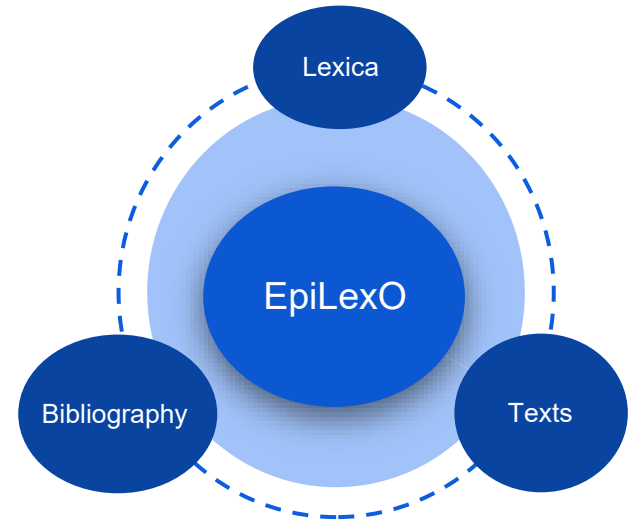
From ItAntDSL to XML-TEI/EpiDoc

- The serialization of ItAntDSL in XML with a proprietary scheme can be easily converted in XML-TEI/EpiDoc through XSL(T) stylesheets
- The conversion is facilitated by the fact that the DSL and XML-EpiDoc share the same abstract model (same structure and same vocabulary)

Bridging Lexica and Inscriptions in a User-Friendly Interface: EpiLexo editor

ItAnt developed a single page, user-friendly application to facilitate scholars in the creation of natively interlinked datasets, centered around lexica and conceptual data.

Currently, EpiDoc scholarly editions of inscriptions are created independently (with ItAnt DSL for Venetic) and ingested as external resources for linking purposes.



EpiLexO editor

a) Navigation Trees (Texts, Lexicon and Lexical Concepts)

b) Editor Area:

- Epigraphy
- Lexicon (pivot)
- Concepts

c) Contextual and additional information

- Attestations
- External linked resources
- Bibliography
- Inscriptions metadata (ingested from Epidoc XML docs)

The screenshot displays the EpiLexO editor interface, which is divided into several panels:

- (a1) Epigraph:** A search bar and a tree view showing a hierarchy of documents: Oscan (ITAnt_Oscan_2.xml, ITAnt_Oscan_7.xml, ITAnt_Oscan_3.xml) and Latin.
- (a2) Lessico:** A search bar and a list of words with their forms and attestations. The word 'upsed' is selected, showing its form (23) and various attestations like (s)itroova, opsens, osins, upsi[id], upsasater, upsatuh, upse[ns], upsed, and uupsens.
- (b1) Epigraphy:** A text editor showing the word 'upsed' in a sentence: 'esidum . prufated . upsed . gaavis . paapiis . gaavieis'. A form is open for 'upsed' with a 'Sense' field and an 'Add a form:' section containing 'upsed' and 'number : singular'.
- (b2) Lexical Entry:** A panel for the lexical entry 'upsed', showing 'DENOTES' and 'COGNATES' sections. A URL is entered in the cognates field: 'https://illa-erc.eu/data/id/lemma/115170'.
- (c1) Attestation Panel:** A panel for the attestation 'ID:2678 - Upsed', showing the URI, author (E. Middel), confidence (1), creator (prova), externalRef (No info), form_id (upsed_verb_osc_upsed_form), and label (upsed).
- (c2) Metadata Panel:** A panel for the metadata of the inscription, showing the Trismegistos ID (TM 170843), other IDs (ST Sa 2, Insk Terzantium 36), and original place (Sammium).

DigitAnt-explore

The screenshot displays the DigitAnt-explore web application interface. At the top, there are navigation tabs: Home, **Inscriptions**, Lexicon, Bibliography, and Concordances. On the right, there are links for EpilExO Search and Advanced Search. The main content area is titled "ItAnt Oscan 2" and "Curse tablet from Monte Vairano". It features two images of the inscription, showing the original and a digital reconstruction. Below the images, there is a "Text Data" section with the following information:

- IDs: [TM 170274](#), [ST Sa 36](#), [Imlt Bouianum 98](#), [Murano 6](#)
- Language: [Oscan](#)
- Alphabet: [Oscan National alphabet](#)
- Word division type: [punctuation](#)
- Inscription type: [defixio](#)
- Ductus: [sinistrorse](#), [face_a](#), [sinistrorse](#), [face_b](#)

On the right side of the interface, there is a "Text" section with a "Text Data" table. The table has two columns: "Face_a" and "Face_b". The "Face_a" column contains the text "1 p[ro]p[ri]e[m] p[ro]p[ri]e[m] p[ro]p[ri]e[m] 1" and the "Face_b" column contains "1 s[er]v[us] s[er]v[us] s[er]v[us]". Below the table, there is a "Translation" section with the text "The curse is in the form of a curse tablet." and a "Support" section with the text "The tablet was found in the area of Monte Vairano, near the village of Monte Vairano, in the territory of the Municipality of Monte Vairano, in the province of Salerno, in the region of Campania, Italy." and a "Material and Preservation" section with the text "Modern find spot name: Monte Vairano (Campania)", "Discovery year: 1979", and "Last known place of conservation: Cambridge".

Integration with CLARIN-IT

DigitAnt software is already open-source and is being described and deposited in the ILC4CLARIN repository

ItAnt-DSL: <http://hdl.handle.net/20.500.11752/ILC-1003>

LexO-server: <http://hdl.handle.net/20.500.11752/ILC-1004>

EpiLexo front-end: <http://hdl.handle.net/20.500.11752/ILC-1005>

Finalised datasets will also be deposited with an open license, and thus described with public and discoverable metadata.

Future Plans

- Workshops, Webinars, and Videotutorials broadcasted through the DiPText Knowledge Centre are planned to share methods and practices related to ItAntDSL
- Improvements of the DigItAnt Web Platform: import/export functionalities, more flexible linking to external resources, ...
- Integration of ItAntDSL as alternative method for creating new scholarly editions

Acknowledgments

“Languages and Cultures of Ancient Italy. Historical Linguistics and Digital Models” is funded by the MUR (PRIN 2017XJLE8J) and hosted by CLARIN-IT.

References

F. Murano, V. Quochi, A. M. Del Grosso, L. Rigobianco, M. Zinzi, *Describing Inscriptions of Ancient Italy. The ItAnt Project and Its Information Encoding Process*, in “Journal on Computing and Cultural Heritage” 16, 3, 2023, pp. 1-14.

Quochi, V., A. Bellandi, F. Khan, M. Mallia, F. Murano, S. Piccini, L. Rigobianco, A. Tommasi, C. Zavattari. 2022. «From Inscriptions to Lexicon and Back: A Platform for Editing and Linking the Languages of Ancient Italy». In *Proceedings of Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022)*, 59–67. Paris: European Language Resources Association.

Quochi V., Bellandi A., Mallia M., Tommasi A. e Zavattari C. (2022) “Supporting Ancient Historical Linguistics and Cultural Studies with EpiLexO”, CLARIN Annual Conference Proceedings 2022, Prague, Czechia, 10-12/10/2022, pp. 39-43.

Any questions?

<https://diptext-kc.clarin-it.it/helpdesk>

Talk to us at our CLARIN Bazaar stall

federico.boschetti@ilc.cnr.it , valeria.quochi@ilc.cnr.it