BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **24**, No 4
Sofia • 2024 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2024-0042

An Interface for Linking Ancient Languages

Michele Mallia, Michela Bandini, Valeria Quochi

Cnr-Istituto di Linguistica Computazionale "A. Zampolli", Via Moruzzi, 1, Pisa, Italy
E-mails: Michele.Mallia@ilc.cnr.it Michela.Bandini@ilc.cnr.it Valeria.Quochi@ilc.cnr.it

Abstract: *This paper focuses on the linking potentials offered by the EpiLexO web-based front-end for creating and editing an ecosystem of digital resources for ancient languages, developed in the context of a project on the languages of fragmentary attestation of ancient Italy. The focus is particularly on mechanisms introduced for linking lexical information to other information bits either internally or externally, e.g., for creating attestations by linking lexical forms to their variants in relevant inscriptions, as well as for linking lexical data to external independent LOD datasets available on a remote endpoint. Finally, in the conclusions, we briefly introduce some future planned or desired enhancements as well as the final platform component, a parallel interface that constitutes the fruition application, which will be open to anyone on the web and will allow for browsing, searching, cross-querying and visualizing the created set of interlinked resources.*

Keywords: *eLexicography, Ancient languages, Linguistic Linked Open Data, Digital historical linguistics.*

1. Introduction

The need to digitally encode both primary and critical data according to well-established, shared formats and practices for web data publication is now appreciated in several cultural heritage disciplines, including epigraphy and historical linguistics. Increasingly more work is now devoted to adapting and developing digital humanities tools to assist scholars in several tasks. Concerning ancient languages and cultures this is particularly true on the side of digital epigraphy, for which several platforms and infrastructures are flourishing for the online publication and sharing of data according to the Linked Data paradigm. Well covered in this respect is the integration and exploitation of commonly shared vocabularies and gazetteers for the encoding, especially of geo-referenced information (e.g., *Pleiades* and *GeoNames*) and archaeological information about the supports of the inscriptions (e.g., the Getty vocabularies, and the EAGLE vocabulary). Recently, also on the side of digitally scholarly editions of ancient inscriptions, efforts are dedicated to developing tools to

assist scholars in their works beyond the TEI/EpiDoc initiative and guidelines (see for instance [1], albeit not directly in compliance to the Linked Open Data paradigm.

One of the most powerful aspects of Linked (Open) Data (LOD) is the possibility of creating interconnected knowledge in the form of federations of semantically interpretable data. The LOD practices make data available to a wider human audience and to machines, thus fostering interdisciplinary collaboration and new discoveries [2, 3]. The inherent power of LOD to refer to other entities of a semantic type, allows us to obtain much more information thanks to the principle of interlinking/networking through the use of relationships and properties between entities defined in ontologies. And, although a great number of lexical resources have been made available in LOD in the last decade, mostly by conversion or retro-digitization, many still lie as islands in the cloud, as they often lack (rich) links to others. Furthermore, conversion always comes with compromises, while LD-native lexicons require a rethinking of the representation of lexical information to best fit graph-based structure [2]. Yet, tools that allow for establishing or revising such ecosystems in a user-friendly way are still missing and the presence of language resources for ancient languages on the Semantic Web is still largely scarce. One possible reason is the complexity of encoding linguistic data according to the LOD best practices and the general lack of this kind of user-friendly tools which could make the endeavor more viable for individual non-digitally-savvy scholars and small research teams. The only notable exception in this respect is the pioneering work represented by the Linking Latin (LiLa) project, which publishes a whole ecosystem of resources for Latin in RDF as LOD [4]. Also, an actionable integration of the different relevant types of information and resources, fundamental for studying ancient cultures through language, is just starting to be experimented with.

In the context of the project *Languages and Cultures of Ancient Italy: Historical Linguistics and Digital Models* (**https://www.prin-italia-antica.unifi.it/**), (ItAnt hereafter), we attempt to address this gap and aim to complement the current digital epigraphy landscape with a user-friendly web platform for creating and then exploring LOD-compliant lexica integrated interlinked, with a coherent set of related resources: critical editions of inscriptions, citations, bibliographic references, and other external available salient (lexical) resources, i.e., LiLa for the time being. The EpiLexO editing application is thus not simply useful for encoding and editing the lexica of ancient languages. It is especially geared to assisting scholars in the (manual) linking of lexical information to other relevant (re-)sources according to the semantic web principles, i.e., its specificity lies in its native capacity to allow resource interlinking, both internally and externally. As the overall system architecture and set-up of the EpiLexO editing application have been described elsewhere [5], here we focus on this latter aspect, under the assumption that the creation of a collaborative web tool with an easy-to-use interface can simplify the work of philologists and historical linguists in the management of lexical and linguistic information about ancient languages, and in facilitating their publication as Linked Data.

The paper is organized as follows. Section 1 presents some related works and the context for the development of the interface and discusses some background. Section 2 briefly describes the data resources used and produced within the platform,

as well as the creation of links. Sections 3 and 4 describe the general architecture of the whole system, as well as the design and development of the EpiLexO front-end interface. In Section 5 we focus on the more interesting and novel linking functionalities of EpiLexO, and we describe them using examples taken from the first nucleus of the Oscan lexicon (the first entries were encoded by Edoardo Middei and Mariarosaria Zinzi at the University of Florence. The lexicon is now complete and available from the Italian CLARIN ILC4CLARIN repository: **http://hdl.handle.net/20.500.11752/OPEN-1023**). Finally, in Section 6 we conclude by sketching some planned future enhancements and introducing the forthcoming exploration interface that will exploit the interlinked datasets created in EpiLexO to perform complex searches and show integrated results.

2. Context and related works

Over the past decades, there has been a significant amount of work on digital epigraphy, resulting in the proliferation of online platforms, each one addressing different languages and/or geographic areas. Particularly relevant and a source of inspiration for our project are: the EAGLE portal (**https://www.eagle-network.eu**), which provides a single access point to the Epigraphic Database of Bari (EDB) (**https://www.edb.uniba.it/**), the Epigraphic Database Heidelberg (EDH) (**https://edh.ub.uni-heidelberg.de/**), and the Epigraphic Database Rome (EDR) (**http://www.edr-edr.it/**); Trismegistos (**https://www.trismegistos.org/**); i.Sicily (**https://isicily.org/**) [6]; Cretan Institutional Inscriptions (**https://ilc4clarin.ilc.cnr.it/cretaninscriptions/en/**) [7], Papiry.info (**https://papyri.info/**), to mention just the most recent and the better-known to Indo-Europeanists. Given their high number, it would be impossible to review all of them and go into the details of their main features within the limited space of this section. Here it will be sufficient to highlight that all of these tools almost exclusively focus on archaeological and historical aspects of the inscriptions; diplomatic and/or interpretative transcriptions are usually provided, but content interactivity and interlinking are generally absent or limited. Also, most of the mentioned portals do not provide online mechanisms for the creation, editing, or annotation of the materials. Many of them include at least related bibliographies, while the most recent ones make use of shared vocabularies for producing standard and reusable metadata descriptors of the items they include. However, most, if not all, of the projects we reviewed lack (interlinked) linguistic information, especially lexical information as usually encoded in (electronic) dictionaries. Our work thus aims at complementing the features of the existing online platforms, with tools for creating and editing Linked Data compliant dictionary data.

Regarding general systems for combining texts, annotations, and lexicons, most available existing solutions are either full-stack applications, which therefore do not satisfy our implementation requirements (see below), or are not capable of handling both XML-based encoding formats and RDF/OWL datasets. EFES (**https://github.com/EpiDoc/EFES**) [8], for instance, is an interesting publishing tool specifically adapted to digital epigraphy, but it is text-centered and does not seem

to straightforwardly allow for integration with external (LOD) lexical datasets. Recogito (**https://recogito.pelagios.org/**), although very interesting and well supported, is again text-centered and seems suited to integrate better gazetteer-like resources than dictionary data. Furthermore, both tools do not offer services such as RESTful APIs, which is instead a strong requirement for our project. Similar considerations hold for systems like INCEpTION (**https://inception-project.github.io/**) [9], a web-based annotation platform that allows for flexible configuration at various levels and a certain degree of knowledge resources integration, but which appears to be particularly fit for entity annotation and linking and also does not expose RESTful APIs.

The application described in this contribution, Epilexo, is developed in the context of a 3-year project on the languages of ancient Italy, which were later supplanted by Latin (e.g., Oscan, Faliscan, Venetic), with the main goal to support the creation of an ecosystem of language resources for ancient fragmentary languages centered on lexicons, in compliance with the current digital humanities and Linked Open Data principles. These languages are also referred to as "languages of fragmentary attestation" because their testimonies comprise a very small number of texts mostly limited to the epigraphic form found in (often severely damaged) inscriptions [10]. Perhaps even more because knowledge about these languages is so precarious (and it is held by a bunch of scholars worldwide), the possibility of digitizing all the available materials and linguistic knowledge is of great importance both for the preservation and availability of such immaterial cultural heritage and for fostering future research (see also [5]). As one of the major project outcomes and also because it addresses specific disciplinary needs, the EpiLexO interface must necessarily first answer project needs and be use-case specific (see Section 3 below); however, its functionalities are general enough to be used for other languages as well (This is still a theoretical possibility, which has not yet been assessed specifically, esp. with languages not targeted in ItAnt. Some different use case application has been proven possible instead at the back-end level, which is designed and developed to be as general and use-case independent as possible (see [11] for details)).

3. Data handled by the platform and reference data models

As anticipated above, the main goal of EpiLexO is to facilitate historical linguists in their daily work of describing and encoding lexicons for archaic languages based on the available primary and secondary sources, and linking lexical data to these sources and possibly other relevant datasets. Therefore, lexicons are the heart of the editing application, which indeed enables scholars to encode multilingual lexical resources enriched with actionable links to their attestations in documented inscriptions, to bibliographical citations, and to other external useful lexical resources, particularly to the LiLa Knowledge Base. Lexical data is encoded online directly via EpiLexO, in compliance with the Ontolex-Lemon lexical model [12] and its extensions. Ontolex is the outcome of an ongoing W3C community effort for the representation and publication of interoperable and actionable linked lexical resources. Although originally developed for adding multilingual lexical data (i.e., part-of-speech,

morphological/inflectional features, form variants, translation equivalents, etc.) to ontological concepts typically contained in domain ontologies [13], it was soon adopted by many projects for modeling different types of lexicons and has become a de-facto standard for the representation and publication of LOD compliant lexical resources, also in the field of digital humanities. Examples of such resources are described for instance in [14-17].

In addition to lexical data, EpiLexO thus handles, when available, digital editions of relevant inscriptions, related bibliography, and other external resources that can be useful to enrich the description and analysis of the target languages. Bibliographic data is not in focus in this paper and will be described in more detail elsewhere (for some more detail one could already [5, 18]). Texts, albeit receiving a visually prominent space in the interface because they are the main primary data on which the study of the languages rests, in EpiLexO have an ancillary role; they are considered instrumental to lexicon encoding, i.e., to link lexical forms to their attestations. Inscriptions are in fact assumed to be encoded independently of the platform, and in ItAnt in fact, new critical editions of selected relevant inscriptions are encoded separately by each expert team according to a common adaptation of the XML TEI/EpiDoc model (**https://epidoc.stoa.org/gl/latest/intro-intro.html**), the de-facto standard for digital epigraphic projects, as described here [19]). Texts can thus be treated as external resources, which are in fact ingested by the platform for their subsequent linking to lexical items, indexing, and searching purposes (as a reviewer acutely pointed out, this may pose problems of persistence and resilience of links. Indeed the platform is based on a strong assumption that implies that the digital editions uploaded should be stable and versioned, and that ensuring that this is actually the case is a user's responsibility. This is a weakness that poses little concern in the native project, and which may be addressed in future enhancements of the platform, or in its integration with other systems specifically geared towards digitally scholarly editions).

In the context of the ItAnt project, linking actually is (mostly) exploited for the representation of etymological information both internally and externally, and of attestations. Etymology is modeled according to the lemonEty model [20, 11] by using a subset of the classes and properties there defined. Linking is useful in our specific case, especially for encoding cognate words (Cognates are defined as words in different languages that share a common ancestor, thus belonging to sibling languages, according to a given etymological reconstruction hypothesis, cfr. [21, 22], e.g., English *father*, German *Vater*, Italian *Padre*, Espaniol *Padre*, French *Père*, …), etymologies and etymons (see Section 5).

4. Design and development of the EpiLexO interface

4.1. Overall DigItAnt architecture

The current implementation of the DigItAnt platform rests on a Service-Oriented Architecture with strong front-end and back-end separation of concerns. It consists of a set of independent software components complying with the OpenAPI

specifications, which provide machine-readable interface files to describe, produce, consume, and display REST services. Each component is devoted to the management of one aspect of the platform, and each one exposes REST APIs that can potentially serve different clients (for details on the overall technical architecture see [5, 18]). The server side consists of two primary back-ends, the LexO-server (for details see [23]) and the CASH-server [24], which manage lexica and textual documents respectively. There is a third server written in NodeJS that handles the transformation of the EpiDoc XML files of the inscriptions according to the Leiden conventions, plus other information such as bibliographic and image data (all software code is available open source: **https://github.com/DigItAnt/**). Both back-ends offer APIs that rely on the HTTP protocol and utilize JSON format for data exchange. To respect the modular nature of the whole system and achieve easy maintainability, we have utilized Docker, a containerization platform that simplifies the deployment and management of the various services within our application. This approach ensures that each component of the platform operates in a self-contained and isolated environment, promoting reliability, scalability, and replicability.

The front-end editing interface in focus in this paper thus operates upon this set of independent software back-end components; it is written in Angular 11, a modern and robust web application framework that facilitates the efficient creation and management of complex user interfaces. To ensure a seamless and visually appealing user experience, it employs Bootstrap, a popular CSS framework for responsive design. The application is designed to facilitate collaboration among scholars, enabling multiple researchers to work simultaneously on the same datasets in cooperation; each user's contributions and actions are tracked and shown so that the platform also fosters a sense of accountability and responsibility. Furthermore, the lexicon editing component provides functionalities for marking the "status" of lexical entries, i.e., in progress, to be revised, and completed with an associated color code. This feature allows users to efficiently manage and monitor the progress of individual entries, while simultaneously maintaining a comprehensive overview of the project's overall status. This is strictly related to user and role management, in such a way that it will be possible to have a group of users who are allowed to encode the data and mark them to be reviewed, and another (group of) user(s) who are responsible for revising and approving the entries, marking them as completed. Notice that the current platform design and development starts in the context of a 3-year national project and should thus be considered a continuously evolving prototype, with new improvements and functionalities added as new or cleared requirements, come up (the work on improving the interface is in fact ongoing. As soon as the back-ends will implement more sophisticated collaborative features, new front-end functionalities may also be integrated).

The development of the whole system followed a mild AGILE method with frequent exchange with our project partners, i.e., historical linguists that represent our typical "end-users", and cyclic testing similar to [25]. Regarding the UI/UX development of the interface, a shallow User-Centered design approach was followed by collecting user requirements at the start of the project also in the form of user stories, creating mock-ups and layout examples, recording users' feedback, and

performing periodic technical evaluation tests on the various developmental stages of the interface so that both back-end and front-end functionalities are periodically tested in terms, especially of reliability and efficiency. Such tests simulate the users' behaviors and assess every functionality focusing on effectiveness and eventual bugs to be fixed or adjustments to be made (to this end, an evaluation grid is shared among team members that report on each feature of the platform. In the grid, we evaluate the efficiency of the various functions, report the presence of bugs, how to reproduce them, what are the actual and expected behaviours, and whether the issue regards the back-ends or the front-end. This approach aligns with the method described in [26] that discusses the importance of frequent reassessment and flexibility in AGILE frameworks to optimize performance and ensure efficient system functioning).

4.2. The interface

EpiLexO is organized into three main sections (or columns), each of them providing a set of different functionalities for different data and information types (for a more detailed description of the interface layout and functionalities see [18]). The column on the left contains the navigation trees for the main resources: corpus, lexicon, and (lexical) concepts; the column on the right side of the interface displays several panels that contain contextual and additional information such as links to related resources, bibliography, attestations, and metadata. The central column is the main working area devoted to lexicon editing and linking operations. It is composed of 2 horizontal panels: the Epigraphy and the Lexicon Editor panels.

The Lexicon Editor is pivotal to the whole platform, modular and contextually adaptive according to the lexical elements selected from the left column. In fact, by selecting a specific item in the navigation tree, the corresponding editing section opens and allows for encoding relevant linguistic properties drawn from shared vocabularies such as *lexinfo* [27]. The Epigraphy panel allows for the creation of external and internal links between texts and lexical items.

Apart from this, linking is mostly available at the lexicon level. Generally linking functions can be of three types, usually at the choice of the users:

i. linking to information encoded within the platform, via querying the back-end while typing in the relevant field;

ii. copy-pasting a URI of a relevant known external resource;

iii. enabling a search on a remote SPARQL endpoint.

The Link panel in the right column offers two generic mechanisms for creating links between any editable element of the lexicon to external relevant LOD resources in the form of SAME AS and SEE ALSO relations, defined in RDF and Ontolex-Lemon, so that linking to external resources can be expressed for any class of the ItAnt lexical model. Furthermore, some of these mechanisms are made available for encoding specific properties based on the specific ItAnt use case and necessities. These specific linking functionalities are one of the core aspects of EpiLexO and are better described in Sections 5.1 and 5.2 below.

5. Linking in action

The following paragraphs illustrate using examples three peculiar functionalities of EpiLexO for interlinking lexical data with internal and external resources, as demanded by the ItAnt use case requirements:

1. linking within the same resource, i.e., for interlinking the lexical items belonging to the lexicons within the platform (Section 5.1);
2. linking to external resources, i.e., for linking lexical items to pieces of information encoded in already existing and external LOD datasets, such as the LiLa Knowledge-base (Section 5.2);
3. linking lexical items to texts, i.e., for representing the attestations of the lexical forms encoded in the lexica (Section 5.3).

5.1. Internal linking within lexicons

One of the most needed features in a lexicon editing tool is indeed allowing for cross-linking between elements of the same (multilingual) lexicon. In Epilexo this can be done in several places and manners. One is the possibility of using the generic mechanisms from the Link panel as anticipated above.

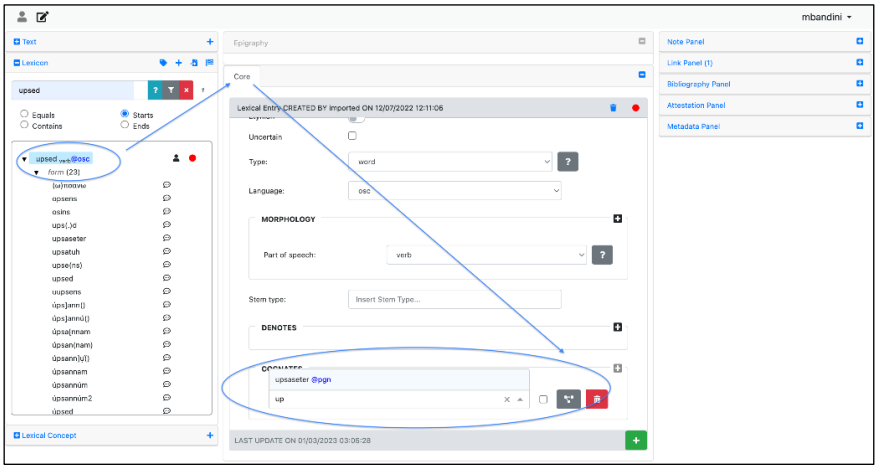

Fig. 1. Internal Linking between two entries for Cognates

Linking options however are available also for different types of lexical information, more specific and peculiar to the Epilexo use cases, particularly for etymological information. For instance, in compliance with the lemon-Ety model [20], the core editing panel for lexical entries allows for the encoding of cognate words. When the cognate to be encoded is already part of the lexicons created within the platform (N.B. the application allows to work on multiple languages), linking is facilitated by triggering a search on the ItAnt lexical dataset directly from the Cognate field. In such a case, the front-end interacts with the LexO back-end for internally retrieving all lexical forms in the other available languages that match the user-typed search string, as shown in Fig. 1, where one can see how the Oscan entry for *upsed* is linked to the entry for *upsaseter* in Paelignan (Paelignan was a language spoken by a tribe that used to live in present-day Valle Peligna, Abruzzo, in the influence area

of Oscan-Sabellic people (**http://www.lexvo.org/page/iso639-3/pgn**)) as its cognate.

Selecting the desired entry triggers a linking action in the underlying back-end and equals to encoding ISCOGNATEOF/HASCOGNATE relations between the two lexical forms. The very same functionality is available for linking to Etymons from within an Etymology, provided that these have previously been encoded as lexical entries in the lexicon.

5.2. Linking to external (lexical) resources

EpiLexO also allows for the creation of links between the lexical items present within the platform and external relevant datasets. In addition to the generic linking mechanisms described above that exploit the RDF SAMEAS/SEEALSO relations, we focus here on the special linking to external LOD resources accessible via remote SPARQL endpoints. By exploiting the LexO-server federation system which permits configuring and sending a predefined SPARQL query to the desired endpoint (for details see [11, p.8]), the front-end interface can hide from the user the complexities of SPARQL syntax and retrieve potential candidate items for linking.

In the specific case of ItAnt, since very few salient resources are available as LOD, this opportunity is exploited in particular for encoding Proto-Indo-European or Proto-Italic etymons and Latin cognates by linking to the LiLa Knowledge Base. EpiLexO thus currently enables linking to the LiLa Lemma Bank [28] and to the Etymological Dictionary of Latin and the other Italic Languages (EDLIL) (**http://hdl.handle.net/20.500.11752/OPEN-533**) [29, 30], with the interface automatically sending contextually salient queries to the LiLa Endpoint (**https://lila-erc.eu/sparql/lila_knowledge_base/sparql**). Fig. 2 below for instance shows the Oscan lexical entry *upsed* being linked to its Latin cognate *opus* from a list of matching items in the LiLa Lemma Bank. Graphically, the query procedure is the same as the one seen in Section 5.1 above for linking internally to encode cognates, so as to improve user experience.

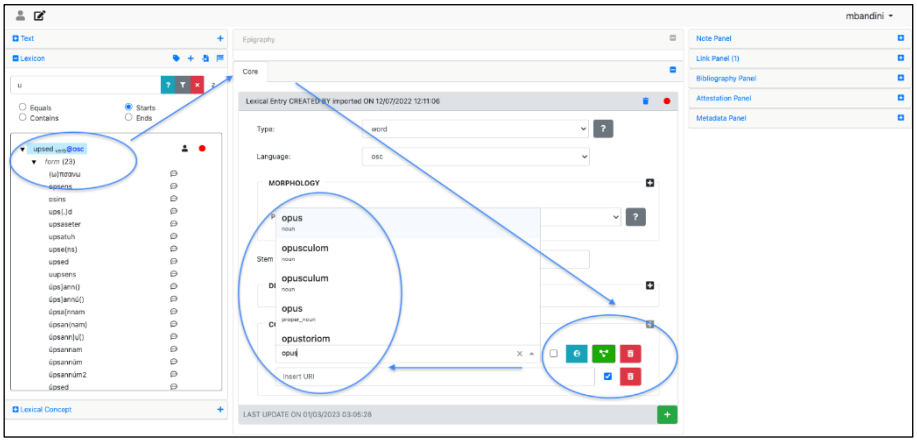

Fig. 2. External linking to a SPARQL endpoint, i.e., LiLa for linking to Latin Cognates

A similar mechanism is available for linking etymological information about a lexical entry to the LiLa Etymological Dictionary of Latin to express the etymons. Fig. 3 below shows how in this case the PIE root **h3ep*, encoded in LiLa, can be selected and linked as the etymon of the Oscan *upsed*.

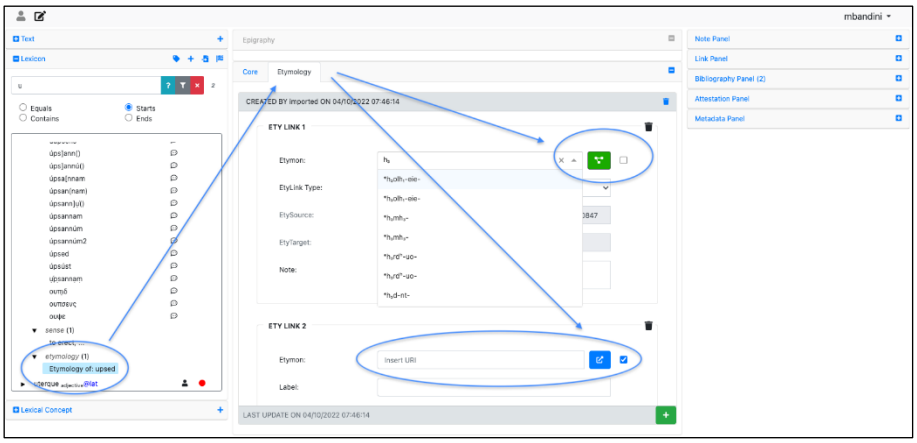

Fig. 3. External linking to LiLa for etymons

5.3. Text-Lexicon linking

Last but not least, an innovative function of the interface proposes to link a lexical form encoded in the internal lexicons to the text(s) in which it is attested, i.e., to the text span in the corresponding digital critical edition contained in the ingested inscription EpiDoc document, when available (as briefly described in Section 3 above). Linking a word occurrence (or variant) in the text with a lexical form in the lexicon, in fact, equals creating an attestation for that form, as shown in Fig. 4 below.

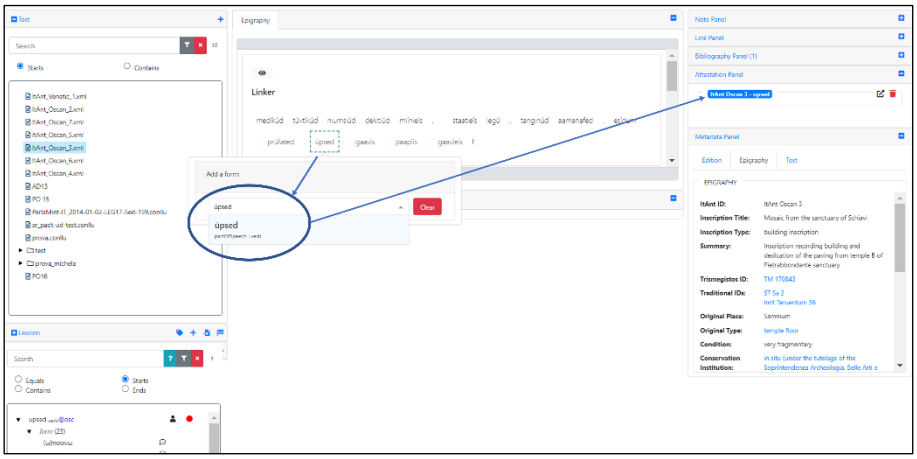

Fig. 4. Text-lexicon linking: create attestations

This operation takes place in the central upper part of the interface, which we may call the Text Linker. Here the text transcription is shown both as a sequence of reconstructed word tokens, which will actually be used for linking, as well as

according to the Leiden transcription conventions, which renders the reconstruction hypothesis operated by the scholar responsible for the specific edition. Attestations, internally, are handled by the CASH server as annotations over text spans and can be enriched, in the Attestation Panel in the right column, with additional information such as relevant bibliography, certainty, authorship, etc.

6. Conclusions and future work

We have presented the recently developed editing application, EpiLexO, created for building interconnected linguistic datasets for ancient and archaic languages, in particular for the languages of ancient Italy. EpiLexO is currently in use within the ItAnt project and its code open sourced (**https://github.com/DigItAnt.** A demo version is online here **https://digitant.ilc.cnr.it/epilexo_demo/ usr: test, pwd:test**. [last accessed 17/10/2024]). This work is based on the belief and assumption that the availability of user-friendly solutions in this discipline can help better support scholars in preserving and sharing their knowledge about those cultures. We have described how the main functionalities of our editing application can be used for creating and interlinking lexical elements with various internal and external resources, focusing especially on the linking capabilities of the application from the interface perspective. Of particular novelty and interest, we claim, are the mechanisms introduced for creating attestations by linking lexical forms to their variants in the imported critical editions represented as spans over the (reconstructed) texts, as well as for linking lexical items and properties to external pre-existing and independent LOD datasets via direct querying remote SPARQL endpoints.

As for future works, from a user experience perspective, we plan to carry out specific user-centered evaluations of usability, including conducting usability tests with a diverse group of users. As suggested by [31], we plan to use some techniques such as forms with targeted questions, addressing a broad audience, to gather feedback on any challenges or areas for improvement regarding the application. This feedback will play a crucial role in refining the interface's design and enhancing its overall user experience. Such an evaluation will include a test on the applicability of the editing tools to different languages, at different historical stages, which is theoretically feasible but not attempted so far, in order also to assess the potential effort needed for adapting, or generalizing, both the interface and back-end services.

One important aspect that has not been previously mentioned, and remains a desired feature in the platform, pertains to the potential for robust customization. It has been observed that allowing users to customize their dataset settings through a dedicated interface, including the ability to modify URIs and namespaces, could greatly enhance the platform's usefulness and usability. By the modular back-end architecture of the system, at least part of the customizations imagined could be easily added in future projects by "simply" adding new back-end services. This can have an impact not just on scalability, but also on the possibility of adding tailored solutions based on evolving user requirements. Additions that we are already thinking of include for instance the possibility to allow users: to choose the vocabulary data properties values; add custom values if needed, or predefine a selection of values

from a given vocabulary to choose from during editing; customize the endpoint(s) to query for external linking. The objective of these enhancements will anyways be to provide a highly adaptive tool that can effectively support the work of historical linguists while remaining intuitive and accessible to users with diverse levels of technical proficiency.

Acknowledging the benefits of a collaborative environment, in future developments we plan to add more features for facilitating teamwork, such as real-time editing and commenting functionalities. This would allow for more efficient and coordinated efforts in the creation and management of linguistic datasets.

Furthermore, a near future addition will be export functionalities for outputting LOD-ready resources. In this respect, lexical data is the least problematic, as it natively conforms to the Ontolex-Lemon model and extensions. For citations, we are investigating existing models such as the FRBR-aligned bibliographic ontology (FaBiO) and the Citation Typing Ontology (CiTO), CFR. [32, 33]; while for attestations we plan to follow the FrAC model, which is a proposed extension to the Ontolex model, albeit still under discussion and revision [34]. Less obvious and straightforward is the conversion of the text documents of the inscriptions. Future work thus includes deepening the investigation of the possibility and advantages of transitioning TEI XML texts to LOD. Currently under scrutiny are some options and models discussed and applied both in recent corpus linguistics and digital humanities for text representation following the LOD paradigms.

One of the options we are taking into account is the NLP Interchange Format (NIF) [35], which however is particularly geared to NLP tools and tasks and thus to represent sentences and their related annotations as LOD; for this format, an opensource conversion tool exists that could be reused or adapted (i.e., CoNLL-RDF [36]) Another option under consideration, closer to the digital humanities feel, is POWLA, a versatile RDF- and OWL/DL-based framework that can be used to represent any kind of linguistically annotated corpus [37-39]. This is in fact the choice made in the LiLa project for their textual resources [40].

Finally, the goal of the ItAnt project is not only to provide a user-friendly editing tool for facilitating the digital LOD-native encoding of ancient lexica with a related ecosystem of linked resources but also important to allow a wider audience to search and study all the available ecosystems of interlinked knowledge, an exploration application is currently under development, which will soon be publicly available.

6.1. The exploration interface

This search and fruiting application, DigItAnt-Search, will be a human-centered access point to the data ecosystem created with Epilexo. It is articulated into two main sections: one focused on the lexicons and lexical access to the ecosystem, the other on the inscriptions/epigraphic texts, and it will guide the exploration based on textual information. In addition, an available advanced query area will permit the user to compose cross-queries over the different datasets. The inscription section focuses on the texts that display all the data about the epigraphic documents encoded in the XML EpiDoc source files and stored in the CASH server. It exploits the contextual metadata coming from the Epidoc headers to build a kind of faceted search based on,

e.g., the subject, inscription type, place of origin, date, and other information. The lexicon section is also divided into various parts according to the different lexical descriptive elements and properties (PoS, Concept, Language, etc.) coming from the LexO-Server.

The interesting feature of this exploration interface is that it will deal with data from several different data sources, not all of which are encoded or ingested through the Epilexo. Specifically, the platform draws external information from several repositories such as Pleiades, LiLa, and GeoNames, supplementing its data with additional insights. One of the features being worked on is the addition of interactivity to the display of the interpretative transcriptions of the inscriptions, i.e., the display of contextually relevant information coming from the other datasets (i.e., lexical information contained in the companion lexica and related bibliographic details and references).

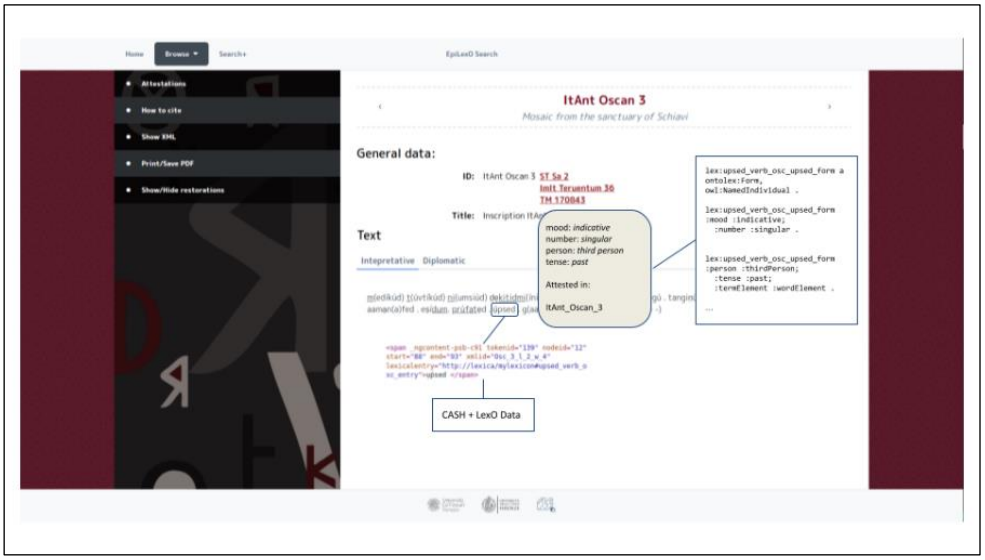

Fig. 5. The exploration application with an overview of a text with a display of text annotations and lexical data provided by the backends

Fig. 5 shows a provisional preview of this explorer. Thanks to the data created by means of the editing application, the exploration interface will allow users (including non-scientists) to view the data produced by scholars, and make advanced searches with the ability to draw on different data sources. The advanced query system will, in fact, search both main servers and cross-reference the data in order to achieve the finest possible search grain.

Also, within DigItAnt-Search it will be possible to visualize the images related to the inscriptions, as indicated in the reference digital scholarly editions ingested. Due to copyright issues for the ItAnt project, only first-hand drawings of the inscriptions will be available. Yet, when available in the XML documents, pointers to sources containing additional images will be shown.

***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 related to the CLARIN-IT research infrastructure.

All authors have made substantial intellectual contributions to this research and have approved the final version of the manuscript. According to the CRediT taxonomy: Valeria Quochi led the conceptualization, provided support in methodology and validation, handled funding acquisition, project administration, and supervision, and contributed to writing and editing the manuscript. Michele Mallia supported the conceptualization, led the methodology, developed the software, and equally contributed to the original draft preparation. Michela Bandini led the validation efforts and equally participated in preparing the original draft.

We thank Andrea Bellandi, Cesare Zavattari, and Alessandro Tommasi for their critical technical contributions to the platform back-end; Emiliano Giovannetti, Simone Marchi, Angelo Mario Del Grosso, and Simone Zenzaro for their valuable discussions and insights; Francesca Murano, Luca Rigobianco, Mariarosaria Zinzi, Carmelina Toscano, Greta Mozzato, and Edoardo Middei for their invaluable contribution as expert users to the discussion and testing of the platform as well as for permission to use their data while still work-in-progress and for their insightful feedback on the GUI applications.

References

1. Baumann, R. SOn of Suda On-Line. – Bulletin of the Institute of Classical Studies, Supplement, 2013, pp. 91-106. **http://www.jstor.org/stable/44216325**
2. Gràcia, J., I. Kernerman, J. B. Gil. Toward Linked Data-Native Dictionaries. – In: Proc. of Electronic Lexicography in the 21st Century: Linking Lexical Data in the Digital Age (eLex 2017) Conference, 2017, pp. 550-559.
3. Bosque-Gil, J., J. Gracia, E. Montiel-Ponsoda, A. Gómez-Pérez. Models to Represent Linguistic Linked Data. – Natural Language Engineering, Vol. **24**, 2018, pp. 811-859. DOI: 10.1017/S1351324918000347.
4. Passarotti, M. C., F. Mambrini. Linking Latin: Interoperable Lexical Resources in the LiLa Project. – In: E. Biagetti, C. Zanchi, S. Luraghi, Eds. Building New Resources for Historical Linguistics. Pavia University Press, 2021, pp. 103-124. **https://hdl.handle.net/10807/194955**
5. Quochi, V., A. Bellandi, M. Mallia, A. Tommasi, C. Zavattari. Supporting Ancient Historical Linguistics and Cultural Studies with EpiLexO. – In: Proc. of CLARIN Annual Conference, 2022, p. 39.
6. Prag, J. R. W., J. Chartrand, 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, de Gruyter Open Poland, 2019, pp. 240-252. DOI: 10.1515/9783110607208-020.
7. Vagionakis, I. Cretan Institutional Inscriptions: A New EpiDoc Database. – Journal of the Text Encoding Initiative (Online), 2021. DOI: 10.4000/jtei.3570.
8. Bodard, G., P. Yordanova. Publication, Testing and Visualization with EFES: A Tool for All Stages of the EpiDoc XML Editing Process. – Studia Universitatis Babeș-Bolyai Digitalia, Vol. **65**, 2020, pp. 17-35. DOI:10.24193/subbdigitalia.2020.1.02.
9. Klie, J.-C., M. Bugert, B. Boullosa, R. E. de Castilho, I. Gurevych. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. – In: Proc. of 27th International Conference on Computational Linguistics (COLING'18), System Demonstrations, Association for Computational Linguistics, 2018, pp. 5-9. **http://tubiblio.ulb.tu-darmstadt.de/106270/**
10. Rigobianco, L. La linguistica delle lingue di attestazione frammentaria. – In: Metodi e prospettive della ricerca linguistica, Vol. **29**, Ledizioni, 2022, pp. 83-94. **https://iris.unive.it/handle/10278/ 3762809**

11. Bellandi, A. Building Linked Lexicography Applications with LexO-Server. – Digital Scholarship in the Humanities, 2023. DOI: 10.1093/llc/fqac095.
12. McCrae, J. P., J. Bosque-Gil, J. Gracia, P. Buitelaar, P. Cimiano. The Ontolex-Lemon Model: Development and Applications. – In: Proc. of eLex 2017 Conference, 2017, pp. 19-21.
13. Declerck, T., P. Buitelaar, T. Wunner, J. McCrae, E. Montiel-Ponsoda, G. Aguado de Cea. Lemon: An Ontology-Lexicon Model for the Multilingual Semantic Web. – In: Proc. of W3C Workshop: The Multilingual Web – Where Are We? Universidad Politécnica de Madrid., Madrid, España, 2010.
http://www.w3.org/International/ multilingualweb/madrid/slides/declerck.pdf
14. Tiberius, C., T. Declerck. A Lemon Model for the ANW Dictionary. – In: I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek, V. Baisa, Eds. Proc. of Conference Electronic Lexicography in the 21st Century: Lexicography from Scratch (eLex 2017) Lexical Computing CZ S.R.O., 2017, pp. 237-251.
15. Abgaz, Y. Using OntoLex-Lemon for Representing and Interlinking Lexicographic Collections of Bavarian Dialects. – In: Proc. of 7th Workshop on Linked Data in Linguistics (LDL'20), European Language Resources Association, Marseille, France, 2020, pp. 61-69.
https://aclanthology. org/2020.ldl-1.9
16. Depuydt, K., J. de Does. Linking the Dictionary of Old Dutch to A Thesaurus of Old English. – Amsterdamer Beiträge zur älteren Germanistik, Vol. **81**, 2021, pp. 493-513. DOI: 10.1163/18756719-12340240.
17. Chiarcos, C., É. Pagé-Perron, I. Khait, N. Schenk, L. Reckling. Towards a Linked Open Data Edition of Sumerian Corpora. – In: Proc. of 11th International Conference on Language Resources and Evaluation (LREC'18), European Language Resources Association (ELRA), Miyazaki, Japan, 2018, pp. 2437-2444.
18. Quochi, V., A. Bellandi, F. Khan, M. Mallia, F. Murano, S. Piccini, L. Rigobianco, A. Tommasi, C. Zavattari. From Inscriptions to Lexicon and Back: A Platform for Editing and Linking the Languages of Ancient Italy. – In: Proc. of 2nd Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA'22), European Language Resources Association (ELRA), 2022, pp. 59-67.
19. Murano, F., V. Quochi, A. M. D. Grosso, L. Rigobianco, M. Zinzi. Describing Inscriptions of Ancient Italy. The ItAnt Project and its Information Encoding Process. – Journal on Computing and Cultural Heritage, Vol. **16**, 2023.
20. Khan, A. F. Towards the Representation of Etymological Data on the Semantic Web. Information, Vol. **9**, 2018. DOI: 10.3390/info9120304.
21. Matthews, P. H. Cognate. – In: The Concise Oxford Dictionary of Linguistics, Oxford University Press, 2014. DOI: 10.1093/acref/9780199675128.001.0001.
22. Campbell, L., M. J. Mixco. A Glossary of Historical Linguistics. Edinbur h University Press, Edinburgh, 2007.
23. Bellandi, A. LexO: An Open-Source System for Managing OntoLex-Lemon Resources. – Language Resources and Evaluation, Vol. **55**, 2021, pp. 1093-1126.
DOI: 10.1007/s10579-021-09546-4.
24. Tommasi, A., C. Zavattari, A. Bellandi, V. Quochi. CASH: A General Purpose Back-End for Corpus Annotation and Search (2024). – In: Vincent Vandeghinste and Thalassia Kontino, Eds. Proc. of CLARIN Annual Conference Proceedings 2024, Barcelona, Spain, 2024.
25. Özkan, D., A. Mishra. Agile Project Management Tools: A Brief Comparative View. – Cybernetics and Information Technologies, Vol. **19**, 2019, No 4, pp. 17-25.
26. Kisimov, V., D. Kabakchieva, A. Naydenov, K. Stefanova. Agile Elastic Desktop Corporate Architecture for Big Data. – Cybernetics and Information Technologies, Vol. **20**, 2020, No 3, pp. 15-31.
27. Cimiano, P., P. Buitelaar, J. McCrae, M. Sintek. LexInfo: A Declarative Model for the Lexicon-Ontology Interface. – SSRN Electronic Journal, 2011. DOI: 10.2139/ssrn.3199505.

28. P a s s a r o t t i, M., F. M a m b r i n i, G. F r a n z i n i, F. M. C e c c h i n i, E. L i t t a, G. M o r e t t i, P. R u f f o l o, R. S p r u g n o l i. Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin. – Studi e Saggi Linguistici, Vol. **LVIII**, 2020, pp. 177-212.
29. de V a a n, M. Etymological Dictionary of Latin and the other Italic Languages. Leiden Indoeuropean Etymological Dictionary Series No 7, Leiden-Boston, Brill, 2008.
30. M a m b r i n i, F., M. C. P a s s a r o t t i. Representing Etymology in the LiLa Knowledge Base of Linguistic Resources for Latin. – In: Proc. of 2020 Globalex Workshop on Linked Lexicography, 2020, pp. 20-28.
31. L i u, F. Usability Evaluation on Websites. – In: Proc. of 9th International Conference on Computer-Aided Industrial Design and Conceptual Design, 2008, pp. 141-144. DOI: 10.1109/CAIDCD.2008.4730538.
32. P e r o n i, S., D. S h o t t o n, F a b i o. CiTO: Ontologies for Describing Bibliographic Resources and Citations. – Journal of Web Semantics, Vol. **17**, 2012, pp. 33-43.
33. D a q u i n o, M., F. G i o v a n n e t t i, F. T o m a s i. Linked data ed edizioni scientifiche digitali. Esperimenti di trasformazione di un Quaderno di appunti. – In: Proc. of AIUCD 2018, 2018, p. 65.
34. C h i a r c o s, C., E.-S. A p o s t o l, B. K a b a s h i, C.-O. T r u i c ă. Modelling Frequency, Attestation, and Corpus-Based Information with OntoLex-FrAC. – In: Proc. of 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 4018-4027.
https://aclanthology.org/2022.coling-1.353
35. H e l l m a n n, S., J. L e h m a n n, S. A u e r, M. B r ü m m e r. Integrating NLP Using Linked Data. – In: C. Salinesi, M. C. Norrie, Ó. Pastor, Eds. Advanced Information Systems Engineering. Lecture Notes in Computer Science, Vol. **7908**. Berlin, Heidelberg, Springer, 2013, pp. 98-113. DOI: 10.1007/978-3-642-41338-4_7.
36. C h i a r c o s, C., C. F ä t h. CoNLL-RDF: Linked Corpora Done in an NLP-Friendly Way. – In: First International Conference Language, Data, and Knowledge (LDK'17), Galway, Ireland, 19-20 June 2017, Proceedings 1, Springer, 2017, pp. 74-88.
37. C h i a r c o s, C. A Generic Formalism to Represent Linguistic Corpora in RDF and OWL/DL. – In: Proc. of 8th International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, 2012, pp. 3205-3212.
http://www.lrec-conf.org/proceedings/lrec2012/ pdf/915_Paper.pdf
38. C h i a r c o s, C. POWLA: Modelling Linguistic Corpora in OWL/DL. – In: Proc. of 9th Extended Semantic Web Conference, The Semantic Web: Research and Applications (ESWC 2012), Heraklion, Crete, Greece, 27-31 May 2012, Proceedings 9, Springer, 2012, pp. 225-239.
39. C h i a r c o s, C. Interoperability of Corpora and Annotations. – Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata, 2012, pp. 161-179.
40. P a s s a r o t t i, M. C., G. P e d o n e s e, R. S p r u g n o l i. Le opere latine di Dante tra annotazione linguistica e web semantico. – Linguistica e Letteratura, Vol. **XLVI**, 2022, pp. 45-71. DOI: 10.5281/zenodo.6514228.

Received: 13.11.2024; Accepted: 20.11.2024. Fast-track.