DigItAnt: a platform for creating, linking and exploiting LOD lexica with heterogeneous resources

Michele Mallia*, Michela Bandini, Andrea Bellandi*, Francesca Murano†, Silvia Piccini*, Luca Rigobianco◇, Alessandro Tommasi*, Cesare Zavattari*, Mariarosaria Zinzi†, Valeria Quochi*

*Istituto di Linguistica Computazionale "A. Zampolli", Consiglio Nazionale delle Ricerche, Pisa, Italy
Area della Ricerca, Pisa, Italy
name.surname@ilc.cnr.it

†Dipartimento di Lettere e Filosofia, Università di Firenze
Firenze, Italy
name.surname@unifi.it

◇ Dipartimento di Studi Umanistici, Università Ca' Foscari
Venezia, Italy
luca.rigobianco@unive.it

Abstract

Over the past few years, the deployment of Linked Open Data (LOD) technologies has witnessed significant advancements across a myriad of sectors, linguistics included. This progression is characterized by an exponential increase in the conversion of resources to adhere to contemporary encoding standards. Such transformations are driven by the objectives outlined in "ecological" methodologies, notably the FAIR data principles, which advocate for the reuse and interoperability of resources. This paper introduces the solutions devised within a nationwide collaborative research project aimed at integrating techniques and methodologies from the conventional study of epigraphic materials, computational lexicography, semantic web, and other digital humanities subfields. It details its services, utilities, and data types and shows how it manages to produce, exploit, and interlink LLOD and non-LLOD datasets in ways that are meaningful to its intended target disciplinary context, i.e. historical linguistics over epigraphic data. The paper also introduces how DigItAnt services and functionalities will contribute to the empowerment of a recently started Italian infrastructure cluster project devoted to the construction of a nationwide federation of research infrastructures for the humanities and cultural heritage, and in particular to its pilot project towards establishing an authoritative LLOD platform.

Keywords: Historical linguistics, Services for linguistics technologies, LLOD, Ontolex-lemon, Digital epigraphy

1. Introduction

The recent years have witnessed a significant technological evolution, accompanied by a parallel methodological development in data processing and utilization. Linguistic technologies, in particular, have seen substantial growth, exemplified by the advancement of expansive linguistic models and tools like ChatGPT. This growth extends to various technological domains within linguistics, including the creation and enhancement of linguistic resources. The increasing adherence to FAIR principles (Wilkinson et al., 2016) and the utilization of Linked Open Data (LOD) (Yu, 2011) have facilitated the emergence of numerous projects, generating valuable resources that have enriched the current data landscape.

Guided by the strategic roadmaps of the European Union and directives from higher institutions, the prevailing policy direction emphasizes data sustainability (European Commission and Directorate-General for Research and Innovation, 2016). The principle here is not to generate data from scratch but to reuse and encode data in a standard format that ensures interoperability for specific applications.

Within the ItAnt project (Marinetti et al., 2021), the DigItAnt platform positions itself within this scientific framework. It aims to establish methodologies and services for creating linguistic resources in LLOD compliant formats for a specific and multidisciplinary area such as digital epigraphy, with a particular focus on historical linguistic aspects.

This initiative, which will be discussed in detail in subsequent paragraphs, is also becoming part of a large infrastructural project named H2IOSC (Humanities and Heritage Italian Open Science Cloud)[1], the ambition of which is to federate all national research nodes into a single entity. DigItAnt's role within H2IOSC is to contribute to piloting the CLARIN-IT LLOD platform by providing a set of web

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

LDL: Linked Data in Linguistics @ LREC-COLING 2024, pages 55–65
20-25 May, 2024.

tools that would allow users to create/update/revise LOD compliant lexical resources (for digital epigraphy) and interlink them with other materials such as digital editions of testimonies, other available LOD lexical and/or conceptual datasets, bibliographic information and common shared vocabularies.

2. Context

This work has been carried out within a 3-year collaborative research project dedicated to expand and advance existing scientific knowledge about the archaic languages of ancient Italy. The *Languages and Cultures of ancient Italy. Historical Linguistics and Digital Models* project (ItAnt henceforth) is thus situated at the crossroad between digital epigraphy and historical linguistics, fields that have experienced significant advancements through numerous interesting projects. In many of these projects, the utilization or publication of linked data is described as presenting opportunities for further growth. However, tools like EFES (Bodard and Yordanova, 2020)[2] and INCEpTION (Klie et al., 2018)[3] facilitate the publication and creation of resources - mostly annotated text corpora - using encoding standards such as TEI-Epidoc (Bodard et al., 2014)[4] or CoNLL, but currently lack the capability to directly produce Linked Open Data (LOD) outputs. Similarly, resource access tools like Institutional Cretan Inscriptions (Vagionakis, 2021) rely on XML technologies like EpiDoc without intending to generate LODified outputs. Some initiatives such as the Epigraphic Database Heidelberg[5] and iSicily[6] (Prag and Chartrand, 2019) recently have leveraged the ability to link data from inscriptions to other data sources (e.g., DbPedia[7]) and have used controlled vocabularies (Pleiades[8], Geonames[9], Trismegistos[10]) for semantically precise and updated metadata annotation (Grieshaber, 2019), but still deliberately do not produce or publish LOD datasets. Within these contexts, spanning epigraphy and other linguistic fields, a need has emerged to tackle one of the most compelling challenges from both a technological and methodological standpoint: to provide (web/virtual) environments enabling scholars to more easily create and access resources available to the humanities public following Open Science paradigms and methodologies promoting interoperability and re-usability. A project related to historical linguistics (specifically to Latin) that fully adheres to LOD standards is the *LiLa: Linking Latin* project (LiLa, for short)[11], which led to the development of various Latin lexical and textual resources, alongside with a suite of tools for analysis, resource linking, and utilization (Passarotti and Mambrini, 2021). Regarding tools, a successful editor for RDF terminological resources is VocBench (Stellato et al., 2015)[12], which has become one of the most comprehensive tools for editing linked data resources in various formats (primarily SKOS, but also Ontolex-lemon (McCrae et al., 2017)), offering collaborative and infrastructural functionalities. The DigItAnt platform positions itself between traditional databases and portals in use in digital epigraphy environments and advanced tools like VocBench.

It represents the first endeavor to integrate functionalities typical of epigraphic databases and web annotation tools into a unified web environment alongside lexicographic tools, facilitating the creation and editing of lexica, vocabularies, and thesauri, as well as to facilitate the interlinking of heterogeneous datasets and publish them as LLOD.

3. DigItAnt Architecture

The DigItAnt platform is developed within the ItAnt project, in collaboration with the Ca' Foscari University of Venice and the University of Florence. Its main goal is to provide scholars with an online environment for creating LOD-ready lexica for the languages of ancient Italy starting from corpora of inscriptions, either already published or autoptically investigated by the project, encoded in TEI-EpiDoc format, and further enrich lexical information by means of linking it to other existing relevant datasets, such as bibliographies and possibly related external lexical resources.

Its service-oriented architecture showcases a dual nature: on the one side, a web application, EpiLexo (Mallia et al., 2023) has been developed to facilitate the editing and accessing of lexical-conceptual data from a triple store (access and manipulation of this data is mediated by a back-end module called LexO-server (Bellandi, 2019), which manages the database triples) and data ingested from XML editions of inscriptions encoded according to the TEI-EpiDoc standard (access and manipulation of this data is handled by the back-end module CASH-server (Zavattari and Tommasi, 2021)). On the other side, a second web application retrieves the data edited and produced by the previous editing interface, making it accessible to users without any need for authentication. To support

[2] https://github.com/EpiDoc/EFES
[3] https://inception-project.github.io/
[4] https://epidoc.stoa.org/
[5] https://edh.ub.uni-heidelberg.de/
[6] www.isicily.org
[7] https://www.dbpedia.org/
[8] https://pleiades.stoa.org/
[9] https://www.geonames.org/
[10] https://www.trismegistos.org/
[11] https://lila-erc.eu/
[12] https://vocbench.uniroma2.it/doc/dev/

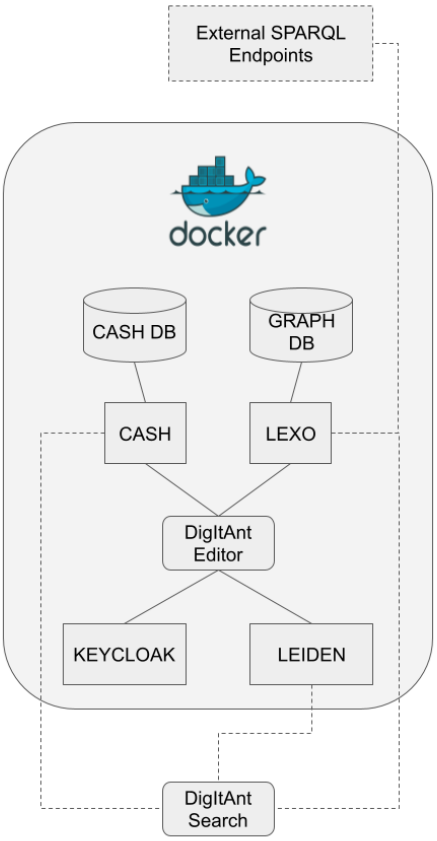

Figure 1: The DigItAnt software architecture

these back-end services, two types of APIs have been prepared: public and private. Only the APIs that allow data retrieval have been made openly available, while editing APIs require an authentication token obtained through user registration[13]. The modular architecture proposed for this project, moreover, as opposed to existing monolithic solutions, potentially allows for various customization, esp. on the front-end side, and improves the application's usability should any of the various back-end services cease to function or become superseded. Furthermore, container technology (specifically, Docker) was chosen to make all applications and services "atomic" and independent from each other. This approach enabled the step-by-step construction of essential components, ranging from the graphical interface to the services for managing LOD data and inscriptions, ultimately leading to the underlying schema (see Figure 1).

In addition, the implementation of authentication via KeyCloak[14] facilitates role mapping among users and makes the platform easily intergratable into federated infrastructural environments. Another important functionality, currently embedded in the LexO-server, is the ability to query external SPARQL endpoints to facilitate linking internal items to external salient resources. The current system offers as a proof-of-concept direct querying to the LiLa endpoint[15] for linking Latin cognate words, etymons and etymologies; specifically, Proto-Indoeuropean and Protp-Italic etyma can be represented by linking directly to the corresponding roots encoded in the *The Etymological Dictionary of Latin and the other Italic Languages in LiLa (EDLIL)* (Mambrini and Passarotti, 2020), while Latin cognates can be linked to the corresponding lemmas in the *LiLa Lemma Bank* (Passarotti et al., 2020). However, the potential to connect with other SPARQL endpoints exists[16].

Beyond this stack lies the exploration and search interface, which makes the data produced with the editing tools accessible in a user-friendly way, and offers a different user experience in comparison to the default SPARQL endpoint, and thus potentially serves different user profiles. In addition to retrieving, filtering and visualizing data from single back-ends of data sources, this interface acts as a kind of middle layer, combining data from different data sources/providers for conducting advanced searches (which, in the current DigItAnt implementation include lexical data in LOD, inscriptions encoded in TEI-EpiDoc, and bibliographic references from Zotero[17]).

Currently, the platform adopts a relatively simple solution for authentication, lacking a genuine federated recognition system. User accounts are custom-created, with rules and authorizations assigned at various levels for resource usage and management. An interface panel facilitates the utilization of these functions, closely integrated with the Keycloak environment. Keycloak possesses the technological capabilities to handle various federated access types through support for multiple secure and legally compliant authentication protocols. Such capabilities should ensure smooth future integration into existing research infrastructures' AAI systems.

Additionally, certain aspects of both front-end interfaces could be improved, particularly regarding the mesh-up and integration of data coming from different heterogeneous sources. For DigItAnt Search, in particular, exploring different data

[13] The exploration interface will be publicly launched and opened upon finalization of the corpus and lexical data at the end of the project (July 2024).

[14] https://www.keycloak.org/

[15] https://lila-erc.eu/sparql/

[16] For more details on the architecture, interfaces and functionalities see also Quochi et al. (2022a) and Quochi et al. (2022b)

[17] https://www.zotero.org/groups/2552746/itant_project/library

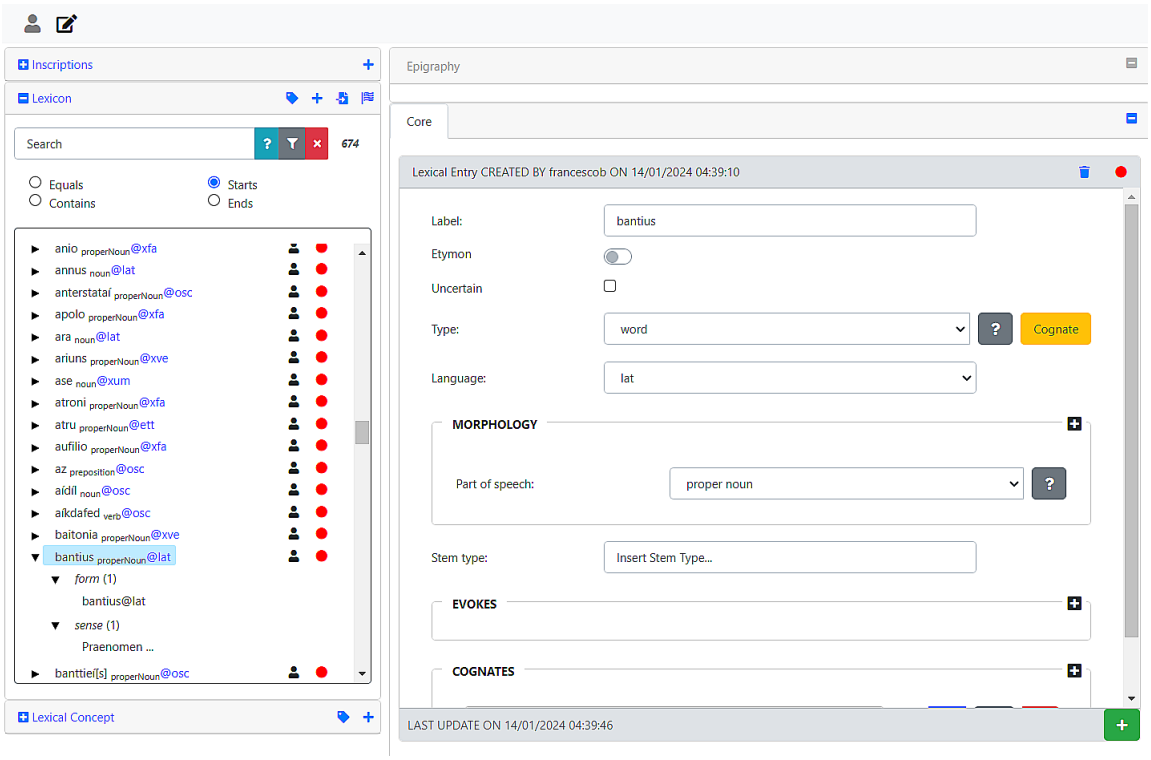

Figure 2: The editing environment

visualization and arrangement possibilities is necessary, depending on the linguistic context in which this service is applied, such as developing tools for information representation based on language or linguistic material type. For the editing platform, an instrumental tool would facilitate managing parameters for LOD material management (e.g., namespace management, workflows, projects, repositories, etc.), making the service accessible to a broader user base.

Finally, it would be advantageous to explore the capability to process a broader typologies of data types and perform multiple computations using certain parameters, such as the selection of metadata types ingested by the server handling texts, and the ability to ingest significantly larger textual corpora compared to the typically small inscriptions.

4. DigItAnt Data

Concerning data models, the platform mainly deals with three heterogeneous data types:

1. lexical data modeled according to the Ontolex-lemon and persisted in a GraphDB instance via the LexO-server;

2. digital scholarly editions of inscriptions encoded according to the TEI-EpiDoc model specifications and ingested from their XML serializations;

3. a bibliographic dataset created and managed via Zotero.

While inscriptions are encoded independently of the editing platform and subsequently ingested as ancillary resources to facilitate the representation of lexica with appropriate attestations, lexica are generated through the platform itself and natively linked to the inscriptions. Additionally, they are linked to relevant bibliographic references via Zotero, and to external lexical-conceptual resources (e.g., through direct queries to the LiLa knowledge-base SPARQL endpoint). Outputs primarily adhere to LLOD formats, with the exception being the ability to export the original XML editions of inscriptions annotated with links to the related lexical forms at the token level. Further details on the EpiDoc customization adopted to address the specificity of the target epigraphical documentation can be found in Murano et al. (2023). Notably, linking with the lexicon is facilitated by the tokenization of each word form in the original XML and the assignment of an @xmlid to each token.

Lexica are at the core of the editing platform. They are designed to be inherently LLOD-ready by adhering to Ontolex-lemon for the model, and LexInfo (Cimiano et al., 2011) for linguistic descriptors, with minimal adjustments to accommodate the special requirements for handling archaic, highly fragmented languages, defined in a project specific ontology.

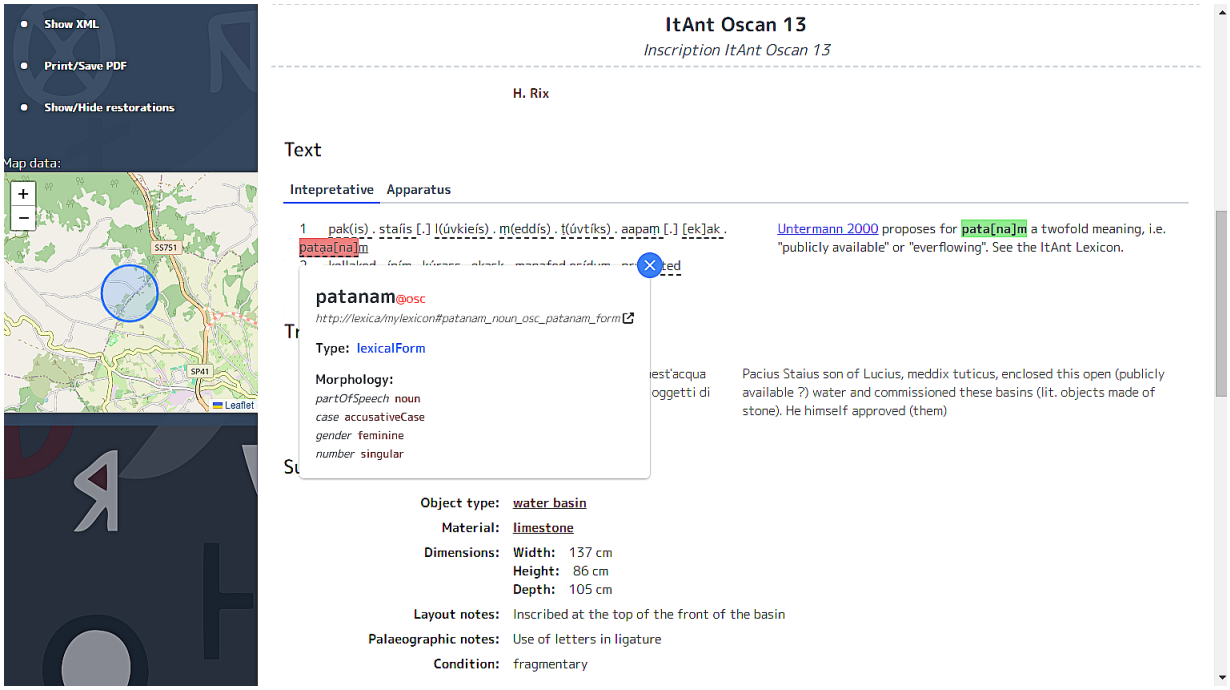

Figure 3: The exploration and search environment

Although digital humanities projects more often adopt TEI XML formats for encoding dictionary data, with TEI Lex-0 becoming a widespread choice, we deliberately chose to model our lexica in Ontolex mainly because: 1. our goal in ItAnt is not to retro-digitize any traditional dictionary, rather to encode the linguistic knowledge that expert scholars formulate on the basis of their interpretation and analysis of the epigraphic texts; 2. for the sake of economy and FAIRness, we wanted to be able to reuse (by linking) available existing (LOD) knowledge; and 3. we wanted to make our outcome actionably available to others. However, because in this project we are dealing with *Restsprachen*, i.e. highly fragmentary attested languages, from ancient Italy –such as Oscan, Faliscan, Venetic, and Cisalpine Celtic– we had to face and find solutions to a number of lexicographic challenges.

First and foremost, because a full paradigm is lacking, it is difficult to retrieve a 'traditional' lemma. Therefore, lexical entries are associated with non-normalized linguistic realization, and no canonical form is formalized. Lexical Entries however still have a label, which is used by the interface for visualization purposes. Due to our limited knowledge of these languages, it is also impossible to provide a thorough description of the syntactic and semantic features typically found in (computational) lexica, such as lexical/syntactic relations or syntactic/semantic roles and frames. From the historical linguistic perspective of ItAnt, etymological information and its level of certainty are instead fundamental. For these reasons, the DigItAnt lexical model uses a subset of the Ontolex Core: i.e. Lexical Entry, Form, Lexical Sense and Lexical Concept; and represents etymological data by exploiting the *lemonEty* extension proposed by Khan (2018) and already used in some important projects, among which the LiLa.

Morphosyntactic representation *Lexical Entry* is the container grouping all the attested forms of a lexical unit. Figure 4 below shows an example[18]. Apart from language and part-of-speech, two additional non-standard data properties are introduced for this class: *stemType*, which roughly indicates noun and adjective classes[19] and *uncertain* for indicating whether the Entry is uncertain.

Form, exemplified in Figure 5, is the key pivotal element of our lexica and encodes standard formal features such as written representation and morphological properties. Word forms in DigItAnt, in fact, correspond to the attested forms, coming from the editor's reading and including the editorial interventions (such as, for example, the restoration of damaged or missing letters). Linking lexical information with the corpus becomes, therefore, fundamental also to ensure reliability. To this end, attestations need to be recorded and encoded for every form, as is usually done in traditional (histori-

[18]The code has been simplified and the URIs have been removed to meet space and template requirements.

[19]For instance, *ā-stems*, i.e. stems ending in *-ā* < PIE *-eh2*, belonging to a specific declension type. LiLa makes use of a similar custom property *inflectionType*.

```
    <!-- Lexical Entry-->
ItAntlex:upsed_entry
   a ontolex:Word;
   rdfs:label "upsed"@osc ;
   lime:language "osc" ;
   lexinfo:partOfSpeech lexinfo:verb ;
   :uncertainty "certain" ;
   ontolex:sense ItAntlex:upsed_sense ;
   ontolex:evokes ItAntlex:toWorkToil_
            semfield_concept .
   ontolex:lexicalForm
           ItAntlex:upsed_opsens_form ;
           ItAntlex:upsed_osins_form ;
           ItAntlex:upsed_upsed_form ;
   lemonEty:etymology ItAntlex:etym_upsed.
    ... .
```

Figure 4: Simplified code snipped of the Lexical Entry for the Oscan verb *upsed*

```
   <!-- Lexical Forms -->
       ...
ItAntlex:upsed_upsed_form
  a ontolex:Form ;
  ontolex:writtenRep "upsed"@osc .
  lexinfo:mood   lexinfo:indicative;
  lexinfo:person lexinfo:thirdPerson;
  lexinfo:number lexinfo:singular;
  lexinfo:tense  lexinfo:past;
  lexinfo:voice  lexinfo:active voice ;
  :cites lexbib:upsed_verb_osc_upsed_
         form_bib583715
    ...                          .
```

Figure 5: A sample of lexical forms encoded in the Entry for the Oscan verb *upsed*

cal) dictionaries. To represent and describe attestations, we plan adopt and adapt the FrAC extension to Ontolex (Chiarcos et al., 2022). Currently, each form of a lexical entry is associated to its exact occurrence(s) in the ItAnt transcribed inscription(s), based on the ingested EpiDoc documents. Attestations are persisted in the CASH-server as text annotations and are enriched with optional information about certainty, authorship, relevant bibliographic citations, and free text notes.

Semantics representation Because for *Restsprachen* it is often not possible to retrieve the accurate semantic content of the words, the provided meanings are mostly generic, and entries generally have one one sense. *Lexical Sense* encoding is therefore minimal; it is specified via a definition, can be indicated as uncertain, and can be associated with a *Lexical Concept*, used in DigItAnt to represent semantic fields. For this purpose, we created a SKOS taxonomy of semantic fields based on Buck's list of semantic fields (Buck, 1949). Among the works concerning the Indo-European semantics, Buck's list is one of the few to have organized the Indo-European lexicon by categories, following a taxonomy[20].

Etymology. As anticipated above, etymology is represented via a subset of classes and properties from *lemonEty*, as exemplified in Figure 6. Etymological information, via *Etymology*, is attached to a *Lexical Entry* and applies to all of its forms. For each lexical entry either or both the Proto-Italic and Proto-Indo-European reconstructed roots are represented and encoded as instances of the class *Etymon*, i.e. Lexical Entries with a special status. Similarly, loanwords may also be reported as such, specifying the relationship with related forms such as *borrowing* rather than *inheritance*. Cognate words attested in sister languages are encoded as instances of another subtype of Lexical Entry established by *lemonEty*, the class *Cognate*. In accordance with the Linked Data principles and so as to avoid to produce data islands, Latin cognates as well as etymons and when deemed relevant Etymologies are linked to the LiLa knowledge base (respectively to the LiLa Lemma Bank (Passarotti et al., 2020)and the EDLIL (Mambrini et al., 2020)

Cognates can be encoded in two ways: 1. by linking externally to another linked data compliant lexicon[21] or 2. by linking internally to a Lexical Entry of a different language, see Figure 7[22].

Finally, bibliographic references and citations of relevant literature can be added/linked to any of the above elements to provide literature regarding the particular lexical information expressed. Currently, *Bibliography* is a system-internal data structure which links directly to the target in the ItAnt Zotero library specifying author, title and date. Furthermore, it makes it possible to specify additional citational information such as page spans and to add free text notes. Ontologies such as CITO (Peroni and Shotton, 2012) are under consideration for exporting citations related to both lexical classes and attestations in the lexicon. Work is also in progress for the mapping of the whole Zotero bib-

[20]The taxonomy, created within the platform, also includes references to the Semantic Index of the Indo-European Lexicon, accessible at https://lrc.la.utexas.edu/lex/semantic, which served as inspiration for our adaptation. It will be disseminated at the end of the project along with the other project outcomes.

[21]This option is viable as regards Latin cognates, for which direct links to a canonical form in the LiLa Lemma Bank can be established directly by means of the *cognate* property. For instance, the Latin cognate of osc. *upsed* 'to erect, to set up, to produce' is represented by the URI of the corresponding lemma in the LiLa knowledge base, namely lat. *opus*.

[22]This option is necessarily used for cognates in languages other than Latin for which LLOD lexica are not available, or when there is no satisfactory match in LiLa.

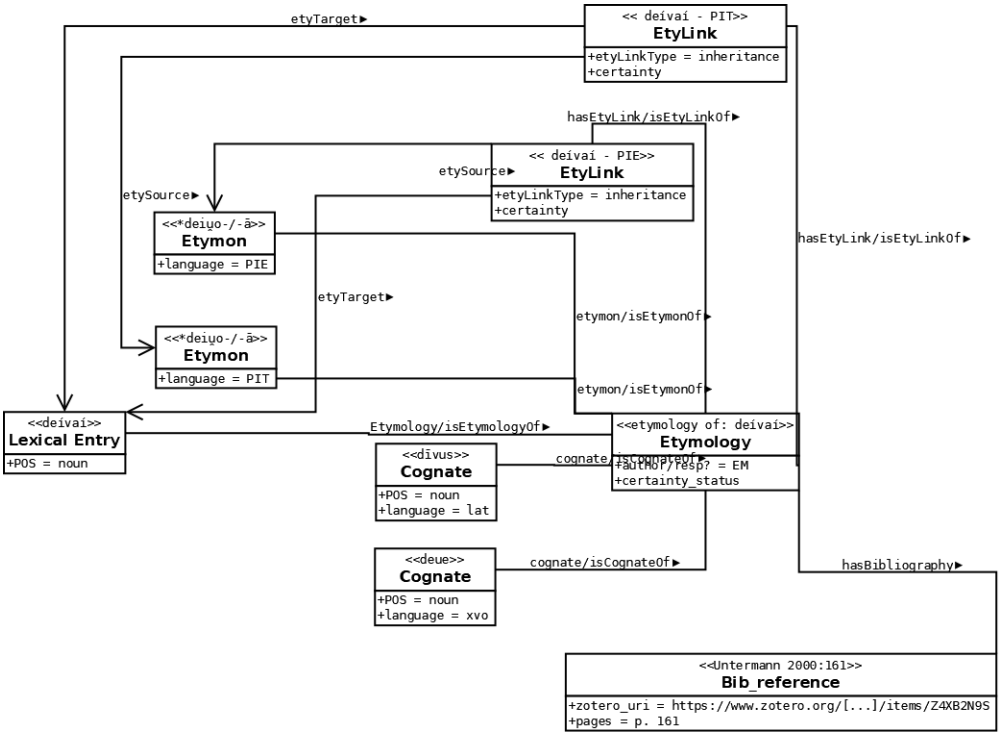

Figure 6: Exemplified model of the etymological information in DigItAnt, according to lemonEty.

```
<!-- Lexical Entry-->
ItAntlex:upsed_entry
    a ontolex:Word;
    rdfs:label "upsed"@osc ;
    lime:language "osc" ;
    lexinfo:partOfSpeech lexinfo:verb;
    lemonEty:cognate lilaLemma:115170;
    lemonEty:cognate ItAntlex:
                        upsaseter_pgn;
```

Figure 7: An example illustrating the two options for encoding cognates. The full URI of *lilaLemma:115170* has been omitted due to space constraints.

liography to FRBRoo/LRMoo[23] (Riva and Žumer, 2018) in order to export and publish the dataset conveniently for versioning and publication.

5. DigItAnt integration into Research Infrastructures

An important objective of the ItAnt project is to contribute to and integrate with research infrastructures for the Humanities, which has partly influenced its technological and data design. Although conceived within a specific project, for a specific use case, the system has been designed to be versatile, potentially benefiting other projects dealing with similar tasks and data. Thus, software components and linguistic data will be distributed and offered as services through the CLARIN-IT infrastructure. This will hopefully enhance support for its community of historical linguists and digital humanists, while also increase the visibility and impact of the project outcomes. The finalized interlinked datasets will be deposited, with open licenses, in the trusted national repository.The source code of the platform components is also released as open-source and deposited for long-term distribution, while the applications themselves (i.e., the front-end applications and the back-end servers) will be distributed through CLARIN channels as docker images for easy installation by any (expert) user.

As the ItAnt project is about to end, platform hosting will also transition to the infrastructure. This transition may serve as a case study within the recently started infrastructural project, Humanities and Heritage Italian Open Science Cloud (H2IOSC)[24], assessing the sustainability and costs of similar scenarios (i.e. short-term funded projects transferring their online outcomes and Virtual Research Environments to the RIs federation).

H2IOSC, in fact, aims to enhance and federate Italian nodes associated with the DARIAH[25], E-

[23] LRMoo is the new ontological bibliographic model developed from FRBRoo. A stable version was released in October 2023.

[24] https://www.h2iosc.cnr.it/

[25] http://dariah.cnr.it/

RIHS[26], CLARIN-IT[27], and OPERAS[28] research infrastructures. Its goal is to provide researchers with wide access to virtual laboratories, data centers and advanced tools for storing, processing, and visualizing digital resources, transcending disciplinary barriers to foster interdisciplinary innovative research.

The collaboration between ItAnt and H2IOSC exemplifies efforts to federate and optimize national research infrastructural resources, incorporating projects that overcome disciplinary boundaries and promote data-driven research in the humanities. This collaboration shall bring mutual benefits to both parties. For ItAnt this partnership ensures the sustainability. Interested scholars will be able not only to explore the project outcomes in the long term, but also to enrich the knowledge (graph) about ancient languages by contributing new data. On the other side, the project serves as a testing ground for H2IOSC's federation solutions and workflows, particularly toward its Linked Open Data (LOD) platform, one of H2IOSC's pilot projects.

DigItAnt may act as a test case for the planned workflows that assist scholars from depositing a (LOD compliant) resource to publishing it in the national endpoint.

6. Conclusion

In this paper we have presented the technological results of a research project that is concluding its activities in July 2024: the current implementation of a platform for creating and exploring linked data about ancient languages and cultures. This platform aims to assist historical linguists in representing their knowledge about these languages and cultures digitally, masking the complexities of dealing with digital models and formats. Centered around lexical data, the unique characteristic of this platform lies in the attempt to mesh-up and interlink heterogeneous datasets. In particular, the platform aims to integrate digital scholarly editions of epigraphic inscriptions, lexical data, citations, bibliographic references, and other relevant external resources. These resources vary not only in type, but also in their representational models and serialization formats (e.g., XML TEI, RDF Ontolex, and Zotero exports). Section 4 briefly described and exemplified their characteristics. The integration and meshing-up of heterogeneous and independent resources are made possible by the underlying Service-Oriented Architecture (SOA), which allows different back-ends to implement suitable technologies for handling various data types and models individually. The orchestration of integrated editing, visualizations, and exports is then delegated to the front-ends and/or middle layers.

The DigItAnt platform will soon be finalized and released as an ItAnt project outcome. It will include export functionalities for the lexicon, attestations, and bibliography, as discussed in Sections 3 and 4, so that the resulting linked datasets may be versioned and deposited in an H2IOSC repository in compliance with the FAIR principles. Within the LOD-platform pilot project, this last event might trigger a procedure that automatically publishes the dataset on the CLARIN-H2IOSC SPARQL endpoint.

In the evolving landscape of digital humanities and cultural heritage research, the integration and optimization of research infrastructures (RI) have emerged as pivotal elements in enhancing interdisciplinary studies and overcoming traditional barriers. Web environments like DigItAnt, which offer sets of web tools for the creation or revision, enrichment, linking, LLOD publication, exploration and search of interconnected digital materials and knowledge about ancient cultures and languages, are good candidates for integration into RIs with mutual benefits. Indeed, an integral component of the H2IOSC vision is the development and refinement of services catering to the diverse needs of the research community. This includes the introduction of novel services. The collaborative paradigm exemplified by the H2IOSC initiative and the integration of projects such as DigItAnt can serve as a model for future developments and integration of data and services into infrastructure clouds. By advocating for a federated approach to research infrastructure, H2IOSC underlines the importance of accessibility, interoperability, and the collective utilization of digital resources, an aspect which will be strengthened by the (L)LOD platform pilot.

Finally, from our list of desired improvements that can further enhance the robustness of the system, we plan to prioritize those that may facilitate the integration into the CLARIN-IT/H2IOSC infrastructure and the LOD pilot. These may include allowing DigItAnt to ingest and manipulate other annotated text formats than TEI EpiDoc, exporting a new version of the original scholarly critical edition of the inscriptions enriched/annotated with the URIs of the lexical items attested, and allowing federated AAI to access the editing functionalities.

7. Acknowledgements

This work is carried out in the context of the PRIN 2017 "Languages and Cultures of Ancient Italy. Historical Linguistics and Digital Models" (no. 2017XJLE8J) funded by the Italian Ministry of University and Research. The DigItAnt platform is also supported by CLARIN-IT. Work on the DigItAnt

[26] https://www.e-rihs.it/
[27] https://www.clarin-it.it/it
[28] https://operas-eu.org/

platform will continue within the H2IOSC Project - Humanities and cultural Heritage Italian Open Science Cloud funded by the European Union NextGenerationEU - National Recovery and Resilience Plan (NRRP) - Mission 4 “Education and Research” Component 2 “From research to business” Investment 3.1 “Fund for the realization of an integrated system of research and innovation infrastructures” Action 3.1.1 “Creation of new research infrastructures strengthening of existing ones and their networking for Scientific Excellence under Horizon Europe” - Project code IR0000029 - CUP B63C22000730005. Implementing Entity CNR.

8. Bibliographical References

Gabriel Bodard, Greta Franzini, Simona Stoyanova, and Charlotte Tupman. 2014. Introducing the epidoc collaborative: TEI XML and tools for encoding classical source texts. In *9th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2014, Lausanne, Switzerland, 8-12 July 2014, Conference Abstracts*. Alliance of Digital Humanities Organizations (ADHO).

Gabriel Bodard and Polina Yordanova. 2020. Publication, Testing and Visualization with EFES: A tool for all stages of the EpiDoc XML editing process. *Studia Universitatis Babeș-Bolyai Digitalia*, 65(1):17–35.

C.D. Buck. 1949. *A Dictionary of Selected Synonyms in the Principal Indo-European Languages: A Contribution to the History of Ideas*. Linguistics/Reference. University of Chicago Press.

Christian Chiarcos, Elena-Simona Apostol, Besim Kabashi, and Ciprian-Octavian Truică. 2022. Modelling Frequency, Attestation, and Corpus-Based Information with OntoLex-FrAC. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4018–4027, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Philipp Cimiano, Paul Buitelaar, John McCrae, and Michael Sintek. 2011. LexInfo: A Declarative Model for the Lexicon-Ontology interface. *SSRN Electronic Journal*.

European Commission and Directorate-General for Research and Innovation. 2016. *Report on the consultation on long term sustainability of research infrastructures*. Publications Office.

Frank Grieshaber. 2019. *Epigraphic Database Heidelberg–Data Reuse Options*. Universitätsbibliothek Heidelberg.

Anas Fahad Khan. 2018. Towards the Representation of Etymological Data on the Semantic Web. *Information*, 9(12).

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018): System Demonstrations*, pages 5–9. Association for Computational Linguistics.

Francesco Mambrini, Flavio Massimiliano Cecchini, Greta Franzini, Eleonora Litta, Marco Carlo Passarotti, and Paolo Ruffolo. 2020. LiLa: Linking Latin Risorse linguistiche per il latino nel Semantic Web (AIUCD 2019). *Umanistica Digitale*, 8.

Anna Marinetti, Francesca Murano, Valeria Quochi, Monica Ballerini, Federico Boschetti, Angelo M. Del Grosso, Silvia Piccini, Luca Rigobianco, and Patrizia Solinas. 2021. Languages and Cultures of Ancient Italy. Historical Linguistics and Digital Models. In *Decimo convegno annuale dell’Associazione per l’Informatica Umanistica e la Cultura Digitale (Pisa, 19 - 22 gennaio 2021)*, pages 528–532, Pisa. Associazione per l’Informatica Umanistica e la Cultura Digitale.

John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The Ontolex-Lemon model: development and applications. In *Proceedings of the eLex 2017 conference*, pages 19–21.

Francesca Murano, Valeria Quochi, Angelo Mario Del Grosso, Luca Rigobianco, and Mariarosaria Zinzi. 2023. Describing Inscriptions of Ancient Italy. The ItAnt Project and Its Information Encoding Process. *Journal on Computing and Cultural Heritage*, 16(3):1–14.

Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. Interlinking through lemmas. The lexical collection of the LiLa Knowledge base of linguistic resources for Latin. *Studi e Saggi Linguistici*, LVIII(1):177–212.

Marco Carlo Passarotti and Francesco Mambrini. 2021. Linking Latin: Interoperable Lexical Resources in the LiLa Project. In Erica Biagetti, Chiara Zanchi, and Silvia Luraghi, editors, *Building new resources for historical linguistics*, pages 103–124. Pavia University Press.

Silvio Peroni and David Shotton. 2012. FaBiO and CiTO: ontologies for describing bibliographic resources and citations. *Journal of Web Semantics*, 17:33–43.

Jonathan R. W. Prag and James Chartrand. 2019. I. Sicily: Building a Digital Corpus of the Inscriptions of Ancient Sicily. In *Crossing Experiences in Digital Epigraphy: From Practice to Discipline*, pages 240–252. De Gruyter Open Poland.

Valeria Quochi, Andrea Bellandi, Fahad Khan, Michele Mallia, Francesca Murano, Silvia Piccini, Luca Rigobianco, Alessandro Tommasi, and Cesare Zavattari. 2022a. 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*, pages 59–67. European Language Resources Association (ELRA).

Valeria Quochi, Andrea Bellandi, Michele Mallia, Alessandro Tommasi, and Cesare Zavattari. 2022b. Supporting Ancient Historical Linguistics and Cultural Studies with EpiLexO. In *CLARIN Annual Conference Proceedings*, page 39.

Pat Riva and Maja Žumer. 2018. FRBRoo, the IFLA Library Reference Model, and Now LRMoo: A Circle of Development. In *Transform Libraries, Transform Societies*, Kuala Lumpur, Malaysia.

Armando Stellato, Sachit Rajbhandari, Andrea Turbati, Manuel Fiorelli, Caterina Caracciolo, Tiziano Lorenzetti, Johannes Keizer, and Maria Teresa Pazienza. 2015. Vocbench: A web application for collaborative development of multilingual thesauri. In *The Semantic Web. Latest Advances and New Domains - 12th European Semantic Web Conference, ESWC 2015, Portoroz, Slovenia, May 31 - June 4, 2015. Proceedings*, volume 9088 of *Lecture Notes in Computer Science*, pages 38–53. Springer.

Irene Vagionakis. 2021. Cretan Institutional Inscriptions: A New EpiDoc Database. *Journal of the Text Encoding Initiative [Online]*.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3.

Liyang Yu. 2011. *A Developer's Guide to the Semantic Web.* Springer.

9. Language Resource References

Andrea Bellandi. 2019. LexO - Lexicographic Editor for Ontolex-lemon resources. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa.

Michele Mallia, Andrea Bellandi, Alessandro Tommasi, Cesare Zavattari, Michela Bandini, and Valeria Quochi. 2023. EpiLexO. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa.

Francesco Mambrini and Marco Passarotti. 2020. The Etymological Dictionary of Latin and the other Italic Languages in LiLa (EDLIL). ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa.

Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. LiLa Lemma Bank - Turtle format. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa.

Cesare Zavattari and Alessandro Tommasi. 2021. CASH - Corpus, Annotation and SearcH. A corpus and annotations management server.

Appendix: a sample entry in turtle format

```
@prefix itant:
</itantproject/ontologies/itant.owl> .
@prefix rdf:
<http://www.w3.org/1999/02/22-rdf-syntax-
ns#> .
@prefix rdfs:
<http://www.w3.org/2000/01/rdf-schema#> .
@prefix ns:
<http://www.w3.org/2003/06/sw-vocab-status/
ns#> .
@prefix ontolex:
<http://www.w3.org/ns/lemon/ontolex#> .
@prefix lime:
<http://www.w3.org/ns/lemon/lime> .
@prefix lilaLemma:
<http://lila-erc.eu/data/id/lemma/> .
@prefix edlil:
<http://lila-erc.eu/data/lexical
Resources/BrillEDL> .
@prefix lemonEty:
<http://lari-datasets.ilc.cnr.it/
```

```
lemonEty> .
@prefix crm:
<http://www.cidoc-crm.org/cidoc-crm/> .
@prefix lexinfo:
<http://www.lexinfo.net/ontology/3.0/
lexinfo> .
@prefix skos:
<http://www.w3.org/2004/02/skos/core> .
@prefix ItAntlex:
</itantproject/data/lexicon#> .
@prefix lexbib: </itantproject/data/
lexicon/bibliography#> .
@prefix semfield: <http://lrc.la.utexas.
edu/lex/ semantic/field/> .

<!-- Lexical Entry-->
ItAntlex:upsed_entry
  a ontolex:Word;
  dct:creator "Edoardo Middei" ;
  dct:contributor "Mariarosaria Zinzi" ;
  ns:term_status "editing";
  rdfs:label "upsed"@osc ;
  lime:language "osc" ;
  lexinfo:partOfSpeech lexinfo:verb ;
  ontolex:sense ItAntlex:upsed_sense ;
  ontolex:evokes ItAntlex:toWorkToil_
                 semfield_concept .
<!-- Etymological info about cognates-->
  lemonEty:cognate lilaLemma:115170 ;
  lemonEty:cognate ItAntlex:upsaseter_pgn ;
<!-- Forms list -->
  ontolex:lexicalForm
       ItAntlex:upsed_opsens_form ;
       ItAntlex:upsed_osins_form ;
       ItAntlex:upsed_upsed_form ;
    ... .
<!-- Lexical Sense -->
ItAntlex:upsed_sense1
  a ontolex:LexicalSense ;
  dct:creator "Edoardo Middei" ;
  skos:definition "to erect, to set up,
  to  produce" ;
  ontolex:lexicalConcept
     ItAntlex:toWorkToil_semfield_concept .
<!-- Lexical Concept -->
ItAntlex:toWorkToil_semfield_concept
  a ontolex:LexicalConcept ;
  owl:sameAs semfield:PA_WV
  (https://lrc.la.utexas.edu/lex/semantic/
  field/PA_WV) .
<!-- Lexical Forms -->
ItAntlex:upsed_opsens_form
  a ontolex:Form ;
  dct:creator "Edoardo Middei" ;
  dct:contributor "Mariarosaria Zinzi" ;
  ontolex:writtenRep "opsens"@osc .
  lexinfo:mood lexinfo:indicative;
  lexinfo:person lexinfo:thirdPerson;
  :cites lexbib:upsed_verb_osc_opsens_
         form_bib682785        .

ItAntlex:upsed_osins_form
  a ontolex:Form ;
  dct:creator "Edoardo Middei" ;
  dct:contributor"Mariarosaria Zinzi" ;
  ontolex:writtenRep "osins"@osc .
  lexinfo:mood lexinfo:subjunctive ;
  lexinfo:person lexinfo:thirdPerson ;
  :cites lexbib:upsed_verb_osc_osins_
         form_bib345190

ItAntlex:upsed_upsed_form
  a ontolex:Form ;
  dct:creator "Edoardo Middei" .
  dct:contributor "Mariarosaria Zinzi" .
  ontolex:writtenRep "upsed"@osc .
  lexinfo:mood lexinfo:indicative;
  lexinfo:person lexinfo:thirdPerson;
  lexinfo:number lexinfo:singular;
  lexinfo:tense lexinfo:past;
  lexinfo:voice lexinfo:active voice ;
  :cites lexbib:upsed_verb_osc_upsed_
         form_bib583715
    ...                          .
<!-- Etymology -->
ItAntlex:upsed_entry
    lemonEty:etymology ItAntlex:etym_upsed .
ItAntlex:etym_upsed
  a lemonEty:Etymology ;
  a crm:E89 ;
  rdfs:label "Etymology of: upsed@osc" ;
  lemonEty:etymon ItAntlex:he3p@PIE_entry ;
  lemonEty:hasEtyLink ItAntlex:etyLupsed-PIE ;
  lexbib:cites lexbib:etymology_412923bib412923 .

ItAntlex:he3p@PIE_entry
  a lemonEty:Etymon ;
  seeAlso edlil:etymon_pie0847 ;
  (https://lila-erc.eu/data/lexicalResources/
  BrillEDL/id/etymon/pie0847)

ItAntlex:etyLupsed-PIE
  a lemonEty:EtyLink ;
  lemonEty:etyLinkType "inheritance" ;
  lemonEty:etySource ItAntlex:he3p@PIE_entry ;
  lemonEty:etyTarget ItAntlex:upsed_entry .
```