REST Services for Corpus management Annotation and SearcH

Alessandro Tommasi
University of Pisa, Italy
ale@ctrl-z-bg.org

Cesare Zavattari
University of Pisa, Italy
cesare.zavattari@gmail.com

Michele Mallia
Cnr-ILC, Pisa, Italy
michele.mallia@ilc.cnr.it

Valeria Quochi
Cnr-ILC, Pisa, Italy
valeria.quochi@ilc.cnr.it

Abstract

This paper presents a back-end software that offers a set of micro web services for the general-purpose management and search of text documents and annotations. Initially developed for a digital epigraphy project, the system focuses on integrating texts and lexicons represented in different paradigms. Nonetheless, the solution is designed to be general and adaptable across various domains.

1 Introduction

The need to digitally encode both primary and critical data in standardised or common formats is widely recognised across several cultural heritage fields, including epigraphy and historical linguistics.

In digital epigraphy, integrating various resources for use by both humans and machines is essential. Unfortunately, this integration remains underdeveloped even in many projects focused on studying ancient cultures through language. Most existing tools and projects, like the Epigraphic Database Heidelberg (EDH)[1] (Grieshaber, 2019) and iSicily[2] (Prag & Chartrand, 2019), primarily focus on the archaeological and historical aspects of inscriptions, but lack the interactivity required to link these resources with linguistic databases. Moreover, these initiatives often lack online, ready-to-use systems for creating, editing, or annotating the digitised materials, hampering collaboration and slowing online availability.

Solutions like EFES (Bodard & Yordanova, 2020) and Recogito (Barker et al., 2019) provide interesting useful tools but are text-centric and lack RESTful APIs, which we believe is crucial for ensuring the versatility of software in modern digital humanities projects.

In the context of ItAnt, an Italian 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[3], we aimed to complement the current landscape by providing a user-friendly web platform, DigItAnt, for creating and exploring LOD-compliant (historical) lexica, natively interlinked with digital editions of inscriptions, and other relevant language resources.

The DigItAnt platform aims to provide a technological solution that meets the needs of historical linguists, whose *modus operandi* includes investigating ancient cultures according to their linguistic documentation. A particular focus is placed on the creation of interlinked linguistic resources according to well-accepted representational models (such as XML TEI and RDF Ontolex-lemon). For details on the overall concept and on the data models adopted see our previous papers (Murano et al., 2023; Quochi, Bellandi, Khan, et al., 2022).

The platform consists of a system of independent software components implemented as a Service-Oriented Architecture. The back-end services are designed to be general, so as to be flexible and serve different use cases. The two main back-ends, the LexO-server and the CASH-server, manage lexicons

[1] https://edh.ub.uni-heidelberg.de/inschrift/suche(last accessed 2024/09/02)

[2] http://sicily.classics.ox.ac.uk/inscriptions/(last accessed 2024/09/02)

[3] *Languages and Cultures of Ancient Italy: Historical Linguistics and Digital Models* (ItAnt), https://www.prin-italia-antica.unifi.it/ (last accessed 2024/09/02)

and (annotated) textual documents, respectively. They both expose APIs based on the HTTP protocol and exchange data in JSON format. A set of other independent components than handle additional functionalities, such as Authentication and Authorisation Information (AAI), handled via an independent Keycloak server instance[4], which can be configured to allow for CLARIN Single Sign On (SSO). Further details about the platform are given in Quochi, Bellandi, Mallia, et al., 2022).

2 The Corpus management and Annotation Component

In this paper, we focus on CASH (**C**orpus, **A**nnotation, and **S**earc**H** server) whose primary responsibility is to serve as the back-end for managing text collections, annotations, and associated metadata. The system was developed to handle richly annotated document collections, including both primary texts and extensive metadata related to their historical and contextual information. Its native use case is to deal with a corpus of *EpiDoc* XML digital critical editions of archaic inscriptions. In addition to the annotated reconstruction of the inscribed texts, the corpus includes a set of contextual, historical, and descriptive metadata, following the practices of digital epigraphy (see Murano et al., 2023 for details on the corpus, and Fig. 4 in Appendix for a simplified sample of ItAnt Epidoc document).

CASH is designed to be modular and extensible in multiple ways, including document ingestion, annotation and metadata semantics, data export, and multi-level queries. The back-end services expose APIs documented via Swagger[5], and the source code is available as open source The source code, written in Java with a MySQL-based persistence layer, is available open source[6].

2.1 Document ingestion

CASH implements a customisable import module that allows users (i.e. project managers that install and set up the back-end systems) to specialise it for handling format-specific requirements. The module already supports three document formats (plain text, CoNLL-X, XML TEI EpiDoc), and is customisable to different models. It also ensures that while the "raw" document data is always available for retrieval, the information can be interpreted during import, creating metadata and annotations according to the specific use case requirements. Metadata can thus be included in the original document (e.g. in a heading statement at the beginning), as in the case for instance of XML TEI or CoNLL-U documents[7].

2.2 Corpus Management and Metadata

CASH organises documents into a file-system-like structure, allowing for easy metadata enrichment, annotation, and efficient searching. Specifically, the management of documents is exposed as a set of Create, Read, Update, and Delete (CRUD) operations, which is a standard practice for systems responsible for managing objects.

In CASH, metadata can be associated both with folders and individual documents, they can be typed and further enriched with additional features, and are available for querying and retrieval, see section 2.4 below. Metadata in fact are represented as a complex objects that have a main key/value structure plus any number of additional sub-features. Interestingly, features (i.e., both keys and values) are not predetermined, so that user clients can choose any to their willing, making the system general and flexible. Metadata values can be scalars, boolean, numerical, string, or composite types, i.e., lists of other key/value types. For example, EpiDoc files encode, among other types, information about the dimensions of the physical support that held the inscription represented in the document. During ingestion, the customised importing module creates a metadata entry that has as key "dimensions" whose value is itself a complex structure composed of a set of key/value pairs that represent all related information (represented by means of curly brackets in the example here below):

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

[5]https://digitant.ilc.cnr.it/cash_demo/swagger-ui/index.html?configUrl=/cash_demo/v3/api-docs/swagger-config add /cash_itant/ before v3/

[6]https://github.com/DigItAnt/CASH-server

[7]While the use of the term "metadata" is arguably improper in computer science terms, in this case, it is still commonly used to convey that the data is not the text itself, but data about it.

```
  ``dimensions": {``precision": ``high", ``unit": ``cm", ``responsible":
DeBenedittis1980ID, ``width": "1.5",``height": "6.7"}
```

whose keys ("precision", "unit", "responsible", "width" and "height") are dictated by the original file encoding scheme. Further interpretation of the metadata and annotations, beyond its importing from the source document format, is left to the user client or to exporting modules. The core system makes no assumption as to metadata and annotation semantics, so as to maintain maximum flexibility. This way the back end may be used to serve different use cases. For instance, it could be used with a front-end specialised for revising or annotating (CoNLL) treebanks.

2.3 Annotation

Like metadata, annotations are represented as complex objects anchored to specific locations in a document's text via their character offsets. For example, part-of-speech tags or word-sense annotations can be attached to tokens or user-defined spans of text.

An annotation is represented with layers and values, which can be largely customised by the user (via the importer or the client). CASH does not impose a fixed set of annotation types, nor it limits the number of attribute-value features an annotation can be enriched with, thus allowing flexibility for different use cases. For instance, they can include Uniform Resource Identifiers (URIs), for facilitating linking operations to external companion resources. Annotations identify spans of unstructured text (potentially multiple spans). At minimum an annotation specifies the `layer`, which defines the type of information provided, and a `value`, i.e., the annotation content. Again, CASH ensures maximum flexibility and does not impose a predefined set of layers or values. For example, in a treebank-like scenario, it would allow a "part-of-speech" layer, with POS tags as possible values (e.g., *NN, VB, ADJ, ...*), whereas in a WSD scenario we might have a "Sense" layer with URIs or IDs identifying senses in an external resource like Wordnet as values. Also, more than one annotation is allowed for the same portion of text. By making the back-end indifferent to specific layers of annotations, as well as to other details, the specificity of the admissible layers and values can be totally determined by the controlling application. The notion of `token` is treated as special kind of annotation, i.e., it is the only layer explicitly known to (and therefore handled specifically by) the system, because it is the most common form of segmentation across many languages and many technological settings[8] Annotations (including tokens) may be further enriched with additional features, which again are entirely determined by the client on the basis of the specific use-case it serves. For all practical purposes, features attached to an annotation can be thought of as a $json$ object, and may be particularly useful in those cases where annotations refer to other annotations.

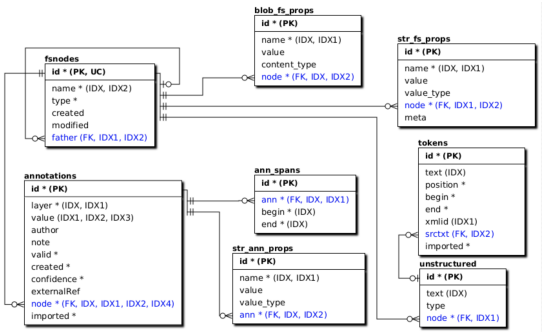

Figure 1: The schema for the underlying database.

Fig. 1 shows a fragment of the database that stores documents along with their associated tokens, metadata, and annotations. The schema includes a table representing "virtual" filesystem nodes (fsnodes), to which the document content, metadata properties (str_fs_props), and an unstructured text stream (the document's content) are linked. The unstructured content can be tokenised, with tokens anchored by

[8] We acknowledge the ongoing debate and criticism surrounding the concept and utility of tokens and tokenisation within the NLP community. However, this is beyond the scope of our discussion here and does not alter the fact that, as of now, tokenisation remains widely used.

their start and end character positions relative to the text. Annotations extend this concept, containing lists of spans (which can be non-consecutive) and associated metadata. Access to the APIs for creating, updating, and deleting documents, metadata, and annotations is secured with OAUTH-issued tokens, managed through a Keycloak installation. APIs for reading and querying the data are instead freely accessible. This way, data may be added and edited only by authorised users, whereas it can be visualised and searched openly.

2.4 Searching

CASH can use all the information persisted in its database for searching the documents. Searching is enabled along three axes: content, metadata, and annotations. User clients can implement functionalities to search for documents containing specific text sequences, metadata fields, or annotations such as tags, or IDs/URIs of standardised information encoded in external companion datasets. These three axes can be combined at will, enabling complex queries. To enhance usability, CASH employs the Corpus Query Language (CQL) (Jakubíček et al., 2010), but extends it to support multi-level searches. CQL has been chosen because it is indeed a well-known and widely used query language in corpus linguistics, likely familiar to many digital humanists, and it already possesses many features we required. Our extension specifically enhance CQL by enabling queries on metadata and supporting sub-token annotations, thus accommodating scenarios where CQL's typical assumption of word separation does not apply.

CASH is designed to be possibly deployed in association with different types of specialised front-end applications and user interfaces, serving different use cases. In practice, so far it is in use within the ItAnt project as one of the back-end serving two fron-end web applications: EpiLexO, for editing archaic lexicons and linking them to related inscriptions (in Fig. 2), and DigItAnt-search, for consuming and querying the DigItAnt data ecosystem (in Fig. 3). The only assumption that the software makes is that a text can be represented as a stream of UTF-8 characters. When this is the case, all other properties of the text (including formatting, typeface, syntactic, semantic layers and so forth) can be represented by annotations anchored to the text by spans over the UTF-8 stream.

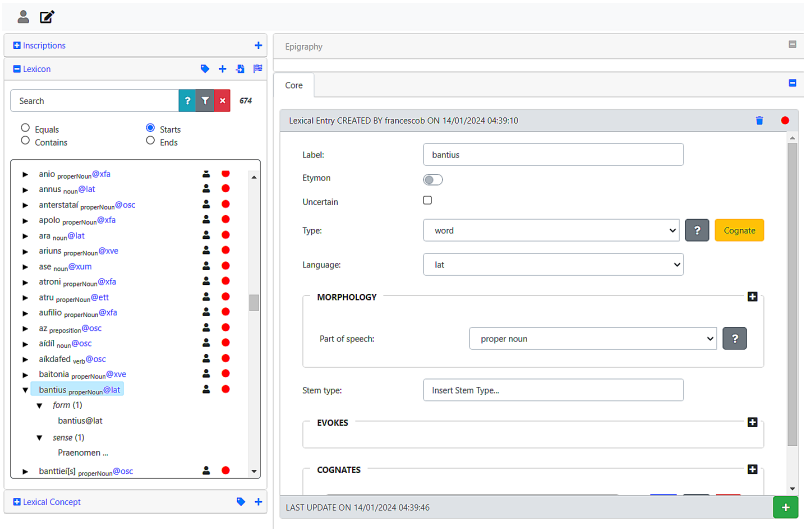

Figure 2: The editing web application

3 Relation with CLARIN-IT

CASH was developed as part of a CLARIN-IT supported project and is one of the key components of the DigItAnt platform[9], a CLARIN-IT/H2IOSC service. The software adheres to FAIR data principles, aligns with Open Science best practices, and is available as open source[10].

[9] https://digitant.ilc.cnr.it/epilexo_search_test/
[10] https://github.com/DigItAnt/CASH-server

Figure 3: The exploration and search web application

4 Acknowledgments

This work was funded by the Italian Ministry of University and Research under the PRIN 2017 program and carried out within the project ”Languages and Cultures of Ancient Italy. Historical Linguistics and Digital Models” (PRIN 2017XJLE8J). It also received technological support from CLARIN-IT.

References

Barker, E., Isaksen, L., Kahn, R., Simon, R., & Vitale, V. (2019). Recogito. https://recogito.pelagios.org/help/about

Bodard, G., & Yordanova, P. (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. https://doi.org/10.24193/subbdigitalia.2020.1.02

Grieshaber, F. (2019). Epigraphic database heidelberg–data reuse options. Universitätsbibliothek Heidelberg. https://doi.org/10.11588/heidok.00026599

Jakubíček, M., Kilgarriff, A., McCarthy, D., & Rychlý, P. (2010). Fast syntactic searching in very large corpora for many languages. *PACLIC*, 741–47.

Murano, F., Quochi, V., Del Grosso, A. M., Rigobianco, L., & Zinzi, M. (2023). Describing Inscriptions of Ancient Italy. The ItAnt Project and Its Information Encoding Process. *Journal on Computing and Cultural Heritage*, *16*(3), 1–14. https://doi.org/10.1145/3606703

Prag, J. R. W., & Chartrand, J. (2019). I. Sicily: Building a Digital Corpus of the Inscriptions of Ancient Sicily. In A. D. Santis & I. Rossi (Eds.), *Crossing Experiences in Digital Epigraphy: From Practice to Discipline* (pp. 240–252). De Gruyter Open Poland. https://doi.org/10.1515/9783110607208-020

Quochi, V., Bellandi, A., Khan, F., Mallia, M., Murano, F., Piccini, S., Rigobianco, L., Tommasi, A., & Zavattari, C. (2022). *Proceedings of Second Workshop on Language Technologies for Historical and Ancient Languages LT4HALA 2022*, 59–67.

Quochi, V., Bellandi, A., Mallia, M., Tommasi, A., & Zavattari, C. (2022). Supporting Ancient Historical Linguistics and Cultural Studies with EpiLexO. *CLARIN Annual Conference Proceedings*, 39.

```
<?xml-model href="https://epidoc.stoa.org/schema/9.4/tei-epidoc.rng"
schematypens="http://relaxng.org/ns/structure/1.0"?>
[...]
<tei:teiHeader>
    <tei:fileDesc>
[...] <tei:editionStmt> <tei:edition> <tei:idno>ItAnt Oscan 2</tei:idno>
</tei:edition>
        <tei:editor> <tei:persName>Francesca Murano</tei:persName> </tei:editor>
      </tei:editionStmt>
      <tei:sourceDesc> <tei:msDesc> <tei:msIdentifier>
        <tei:settlement ref="https://sws.geonames.org/3180991">Campobasso
        </tei:settlement>
[...]
        <tei:institution ana="<url>">Soprintendenza Archeologia, Belle Arti e
        Paesaggio
            del Molise</tei:institution>
            <tei:idno>3974</tei:idno> </tei:altIdentifier>
            <tei:msName>Curse tablet from Monte Vairano</tei:msName>
            <tei:altIdentifier type="trismegistos">
                <tei:idno source="www.trismegistos.org/text/170774">TM 170774
                </tei:idno>
    [...]   </tei:altIdentifier> </tei:msIdentifier>
[...]
    <tei:physDesc>
     <tei:objectDesc> <tei:supportDesc> <tei:support>
        <tei:objectType ana="http://vocab.getty.edu/page/aat/300223016">tablet
        </tei:objectType>
[...]
        <tei:dimensions type="objectDimensions" unit="cm" precision="high"
           resp="#De_Benedittis1980a">
            <tei:height quantity="6.7">6,7</tei:height>
            <tei:width quantity="1.5">1,5</tei:width></tei:dimensions>
[...]   </tei:supportDesc>
        <tei:layoutDesc>
            <tei:layout columns="2" writtenLines="4">
            <tei:rs type="execution"
            ana="http://vocab.getty.edu/page/aat/300404794">
            engraved</tei:rs>
[...]   </tei:layout> </tei:layoutDesc>
     </tei:objectDesc>
     <tei:scriptDesc>
        <tei:scriptNote>
            <tei:rs type="writingSystem" subtype="alphabet" ref="#oscan-etruscan">
            Oscan National alphabet</tei:rs>
            <tei:rs type="wordDivision">punctuation</tei:rs>
            </tei:scriptNote> </tei:scriptDesc>
    </tei:physDesc>
    <tei:history> <tei:origin>
        <tei:origPlace type="composed">
          <tei:placeName type="ancient"
          ref="https://pleiades.stoa.org/places/438681">
          Aquilonia, Samnium</tei:placeName>
          <tei:placeName type="modern" ref="https://sws.geonames.org/3164966">
          Monte Vairano (Campobasso)</tei:placeName>
        </tei:origPlace>
 [...]
    </tei:history> </tei:msDesc> </tei:sourceDesc> </tei:fileDesc>
 [...]
    <tei:text>
     <tei:body>
        <tei:div type="edition" subtype="interpretative" xml:space="preserve">
          <tei:div type="textpart" n="face_a" style="text-direction:r-to-l"
          rend="ductus:sinistrorse">
            <tei:ab>
              <tei:lb n="1" xml:lang="osc-Ital-x-oscetr" xml:id="Osc_2_1_1"/>
[...]
                <tei:name type="patronymic" xml:lang="osc-Ital-x-oscetr"
                xml:id="Osc_2_1_1_w_3" ref="#p1">
                <tei:expan><tei:abbr>tre</tei:abbr>
                <tei:ex>bieís</tei:ex></tei:expan></tei:name>
            </tei:ab>
[...] <tei:div type="translation" xml:lang="eng">
            <tei:p>Pacius Helvius son of Trebius | Statius Betitius [son of ...]
            </tei:p>
[...] </tei:div>
      <tei:div type="commentary" xml:lang="eng" resp="Francesca Murano">
 [...]</tei:div>
      <tei:div type="bibliography">
 [...]</tei:div>   </tei:body> </tei:text>
```

Figure 4: A sample of an ItAnt EpiDoc document. Most encoded parts are omitted for reasons of space. This serves simply as an exemple of the richness of information handled.