# REST Services for Corpus management Annotation and SearcH

Alessandro Tommasi & Cesare Zavattari[1], Michele Mallia & Valeria Quochi[2]
[1]University of Pisa, [2]CNR-ILC,
[1]name@ctrl-z-bg.org, [2]name.surname@ilc.cnr.it

CLARIN-IT

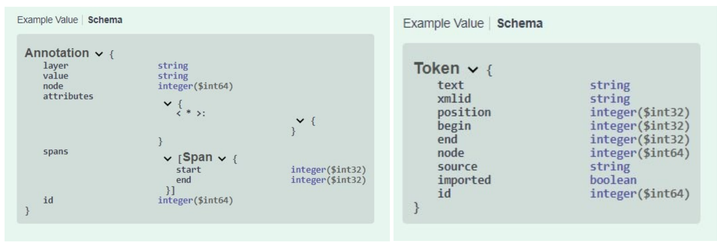
CLARIN
Common Language Resources and Technology Infrastructure

## CASH (Corpus, Annotation, and SearcH) design

A back-end for managing text collections, annotations, and associated metadata. Developed to handle richly annotated documents, including extensive (contextual) metadata. It is modular and extensible in multiple ways, including document ingestion, annotation and metadata semantics, data export, and multi-level queries. The source code is written in Java with a MySQL-based persistence layer. Natively able to ingest EpiDoc XML, CoNLL-X, txt plain documents, it can be customized for various types of input formats.

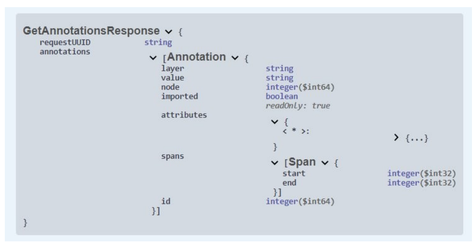
## Text Management, Metadata, Annotation

- CRUD operations
- Metadata: associated both with folders and documents; are complex objects with a main key/value structure plus additional not predetermined sub-features; are available for querying and retrieval
  e.g. ``dimensions": {``precision": ``high", ``unit": ``cm", ``responsible":DeBenedittis1980ID, ``width":"1.5",``height": "6.7"}
- Annotations: complex objects anchored to specific locations via character offsets; represented with layers and values; no imposed fixed set of annotation types, nor number of attribute-value features per annotation;
  → interpretation of the metadata and annotations is clients responsibility.



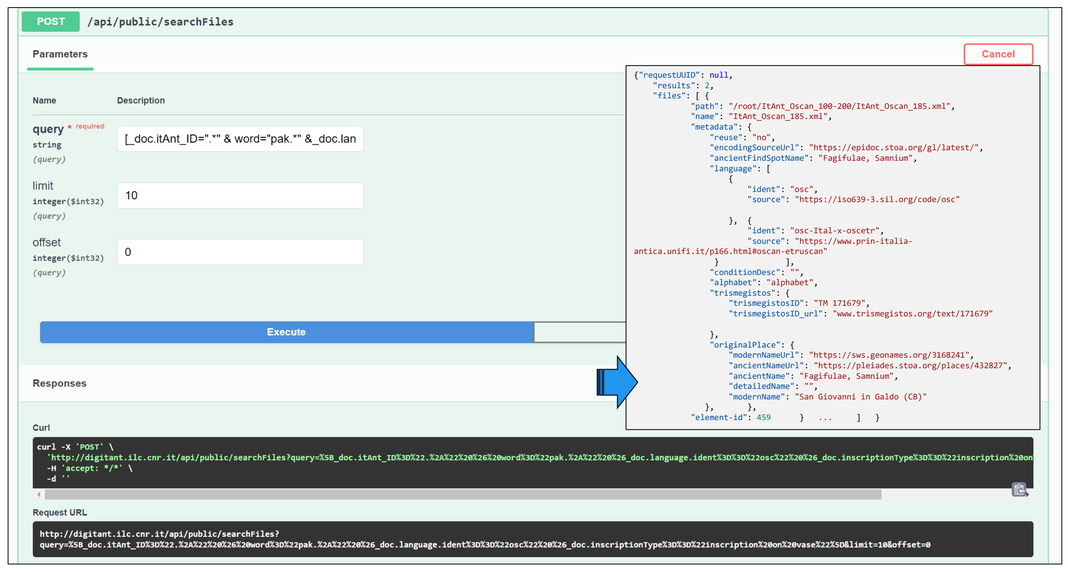






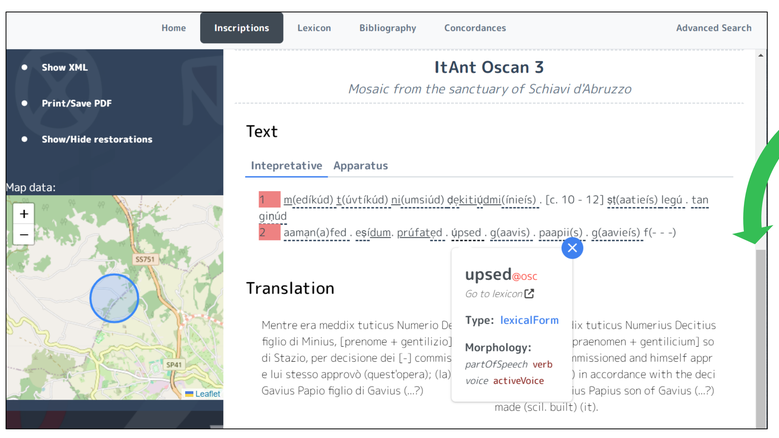

## Query & Search

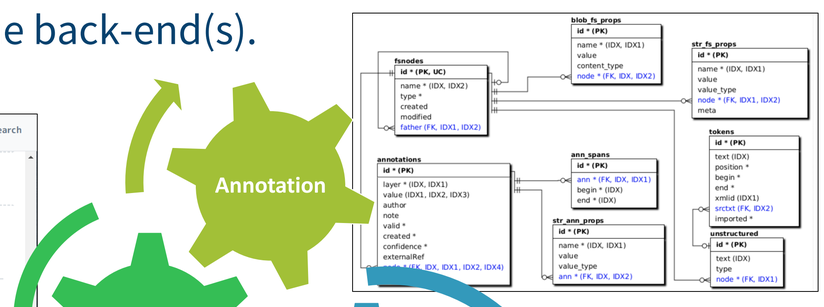
- all info persisted can be object of queries;
- three axes: content, metadata, and annotations; these can be combined at will, enabling complex queries;
- extends Corpus Query Language (CQL) to support multi-level searches. → enables queries on metadata and supporting sub-token annotations, overcoming CQL's typical assumption of word separation.




# CASH for Digital Epigraphy and historical linguistics: the DigItAnt use case

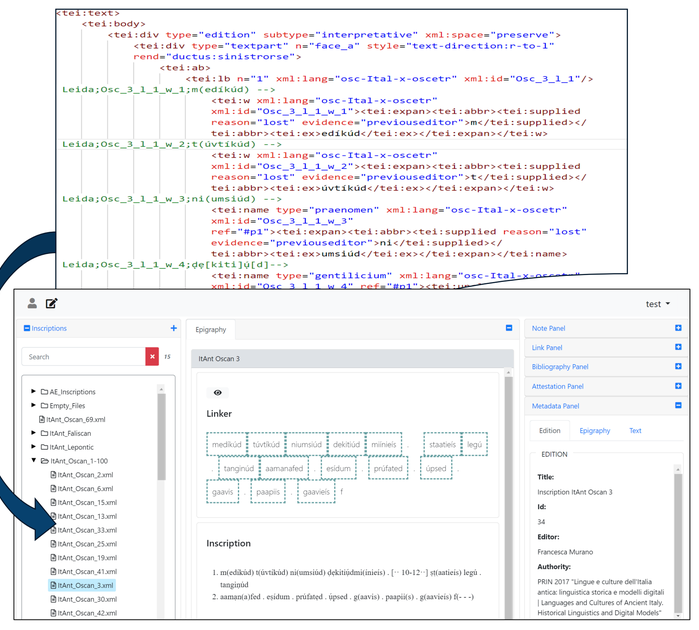
Two distinct front end applications, same back-end(s).













A European Research Infrastructure

CLARIN ERIC was established in 2012 and received ESFRI Landmark status in 2016

## www.clarin.eu

https://github.com/DigItAnt/CASH-server

https://digitant.ilc.cnr.it/platform

clarin@clarin.eu

github.com/clarin-eric

## Acknowledgements

Project: "Languages and Cultures of Ancient Italy. Historical Linguistics and Digital Models", funded by the Italian MUR under the National Strategic Research Grant (PRIN 2017XJLE8J) . The work was also supported and related to CLARIN-IT.