A cloud-native web application for AI-powered queries using fine-tuning on pre-trained Natural Language Models for Named Entities Recognition - THESPIAN-NER

Authors: Alessandro Bombini[1], Lisa Castelli[1], Achille Felicetti[2], Franco Niccolucci[2], Anna Reccia[2], and Francesco Taccetti[1].

1.Istituto Nazionale di Fisica Nucleare (INFN), Sezione di Firenze, Florence, Italy; **2.**Vast-Lab, PIN, Prato, Italy

Abstract

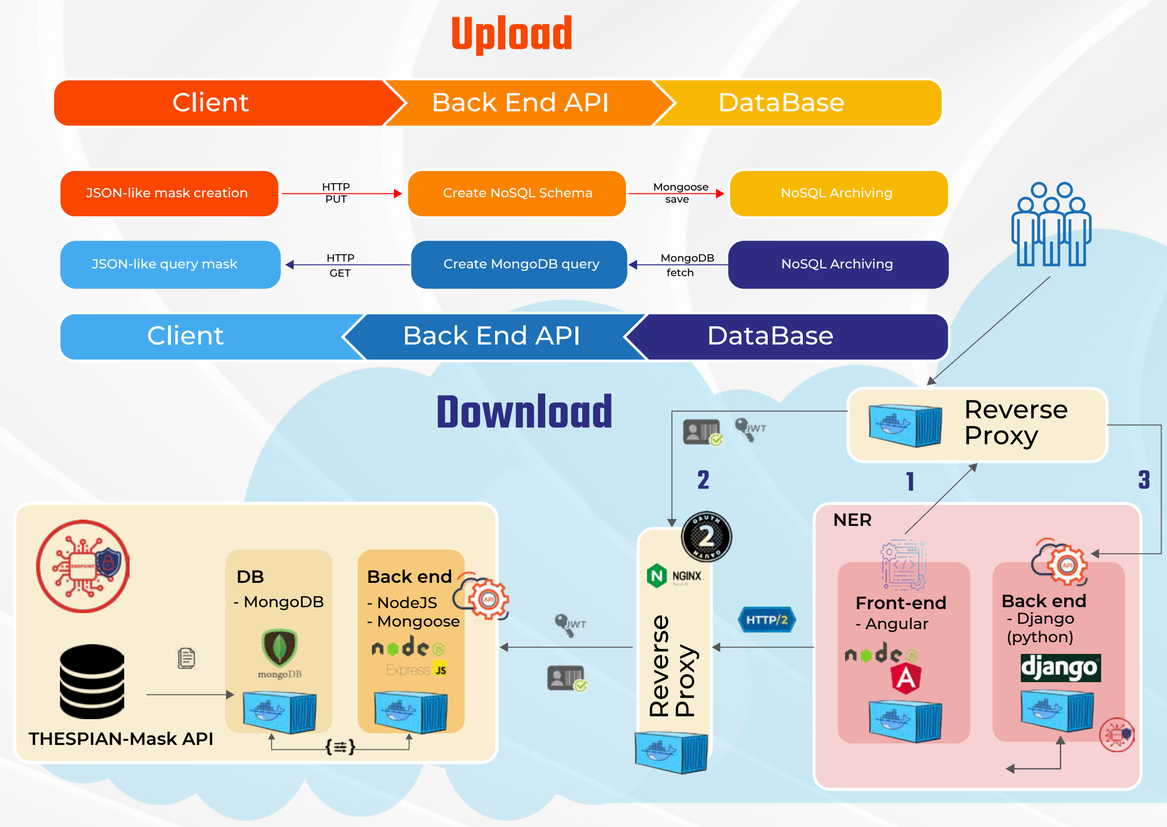
When researchers and professionals try to retrieve stored analysis raw data, they usually face three major issues: data availability, missing (meta)data standardisation, and difficult query creation for database fetching.

To ease these issues, in the framework of the Eosc-Pillar project, the cloud platform *Tools for HEritage Science Processing, Integration, and ANalysis* (**THESPIAN**) was developed [1].

The first web service of the THESPIAN platform, THESPIAN-Mask, was developed to resolve the first two issues; it is a web-app for FAIR storage of scientific analysis on Cultural Heritage, and it is based on an hoc developed CIDOC-CRM-compatible ontology, CRMhs, describing the metadata of scientific files.

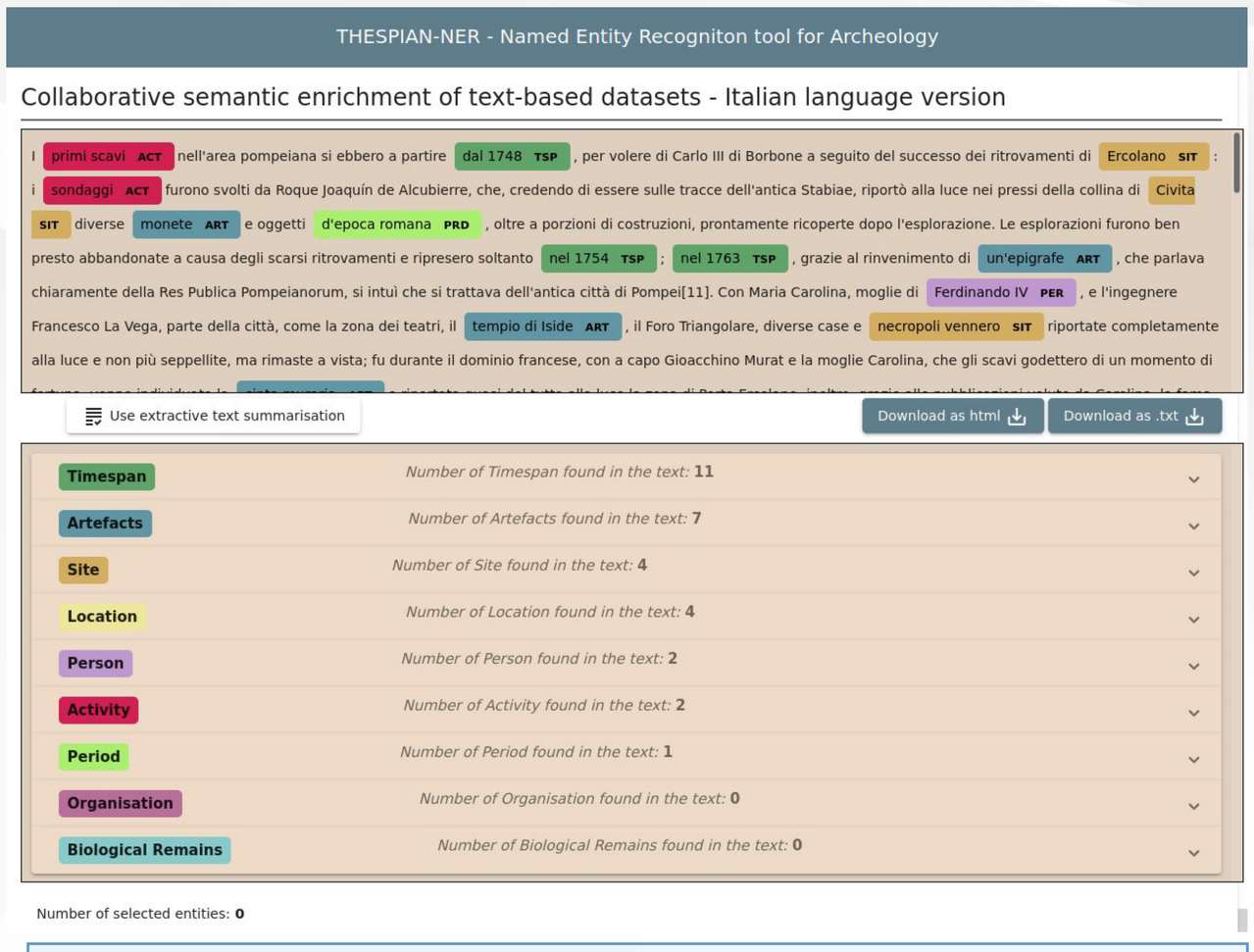
To ease the third issue, and also to help the metadata generation process, an additional cloud-native tool was developed: **THESPIAN-NER**.

It is a tool based on a deep neural network for Named Entity Recognition, enabling users to upload their Italian-written report files and obtain labelled named entities usable as keywords, either for (semi)automatically compose custom queries to the database, or fill (part of) the metadata form describing the file to be uploaded.



Why THESPIAN-Mask?

- ★ Implements FAIR principles [2]
- ★ Offers assisted metadata generation
- ★ Cloud storage
- ★ Based on ad-hoc ontology, makes information fully interoperable with other data



THESPIAN-NER web-app

Front-End Web UI for

- ★ Upload .txt, .pdf documents
- ★ Easily compose custom queries
- ★ Easily fill in the metadata form

Back-End RESTful-API(s) for

- ★ Perform NER anaysis of uploaded text
- ★ Query the DataBase
- ★ Open THESPIAN-Mask

The Authentication/Authorisation process is handled by the reverse proxy

Why THESPIAN-NER?

- Offers assisted metadata generation
- Based on ad-hoc ontology, makes information fully interoperable with other data

To ease these two issues, the cloud-native web-app THESPIAN-NER was developed

- (semi)automatically generate custom THESPIAN-Mask queries, by similarity with a user-selected italian written documents or reports;
- (semi)automatically fill-in a certain number of THESPIAN-Mask metadata fields, again by similarity with a set of Italian written documents or reports;

DATASET(s)

ArcheoNER:

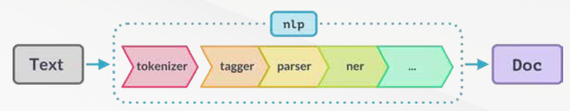
- ★ 9 Entity labels
- ★ 92 Italian-written docs
- ★ 5230 total entities

hsNER

- ★ 14 Entity labels
- ★ 43 italian-written reports
- ★ 5676 entities

Fine tune the spaCy NER model - I

- ★ open-source models used;
- ★ two models used:
  - ☆ CNN-based
  - ☆ Transformer-based



Fine tune the spaCy NER model - II

- ★ Two training for each model:
  - ☆ ArcheoNER, for Archaeological documents
  - ☆ hsNER, for internal lab reports on heritage science.

	ArcheoNER	hsNER	THESPIAN-Mask JSON keys
ACT	✓	✓	sample.preparation.method
ANL		✓	analysis.category.type
ART	✓	✓	studyObject.name
BIO	✓		studyObject.name
DEV		✓	analysis.device
LOC	✓	✓	studyObject.locationLabel
MAT		✓	sample.material
MET		✓	sample.preparation.method
NAT		✓	studyObject.name
ORG	✓	✓	studyObject.owner
PER	✓	✓	studyObject.author
PRD	✓		studyObject.periodLabel
SAM		✓	sample.type
SIT	✓		studyObject.provenanceLabel
SOF		✓	analysis.device.software
RES		✓	analysis.result
TSP	✓	✓	analysis.startDate(endDate)

Cons:

- Difficult to get huge, high quality training data;

Pros:

- Model still get the most relevant entities for the tasks (metadata generation and query creation)

Acknowledgments

The present research has been partially funded by the European Commission within the Framework Programme Horizon 2020 with the projects EOSC-Pillar (Grant agreement number H2020-INFRAEOSC-05-2018-2019-857650)