DOI: 10.26314/GARR-Conf21-proceedings-09

CHNet cloud: an EOSC-based cloud for physical technologies applied to cultural heritages

Alessandro Bombini[1], Lisa Castelli[1], Luca dell'Agnello[2], Achille Felicetti[3], Francesco Giacomini[2], Franco Niccolucci[3], Francesco Taccetti[1]

[1]Cultural Heritage Network, Istituto Nazionale di Fisica Nucleare, Florence. [2]CNAF, Istituto Nazionale di Fisica Nucleare, Bologna. [3]VAST-LAB, PIN, University of Florence

Abstract. In the framework of the European projects ARIADNEplus and EOSC-Pillar, a cloud system, named Tools for HEritage Science Processing, Integration, and ANalysis (THESPIAN), was developed, which offers multiple microservices to the researchers of the Cultural Heritage Network (CHNet) of INFN (Istituto Nazionale di Fisica Nucleare), from storing their raw data to reuse them by following the FAIR principles for establishing integration and interoperability among shared information. The CHNet cloud currently offers three web services: THESPIAN-Mask, a service for assisted metadata generation and data storage, based on an ad-hoc developed ontology called CRMhs; THESPIAN-NER, a tool based on a deep neural network for Named Entity Recognition, which can interpret Italian-written archeological documents and annotate them extracting named entities, that are used for devising custom CHNet database queries; and XRF analyser, a tool for on-line, real-time elaboration of raw data of X-Ray Fluorescence imaging analysis performed on Cultural Heritage.

Keywords. cloud services, FAIR Data management, Big data and Smart Data, AI Machine and Deep Learning.

Introduction

In the framework of the European projects ARIADNEplus and EOSC-Pillar initiative of the European Open Science Cloud (EOSC) framework, Tools for HEritage Science Processing, Integration, and ANalysis (THESPIAN) was developed. THESPIAN is a cloud system offering multiple microservices to the researchers of the internal INFN network devoted to the application of physical technologies to Cultural Heritage, the so-called Cultural Heritage Network (CHNet), and to all the external researchers who cooperate with the network. The mission of CHNet is to harmonise and enhance the expertise of the Institute in the field of Cultural Heritage, expertise distributed among many structures spread over the whole Italian territory. CHNet includes the INFN research groups whose activity is devoted to the development and application of technologies for the study and conservation of Cultural Heritage, but it is open also to other national partners with expertise complementary to that of the Institute and to international partners engaged in diagnostics in cultural heritage. At the moment, the network is composed of three layers:

1. first level nodes: INFN structures;
2. second level nodes: Partners with complementary competencies, such as restoration

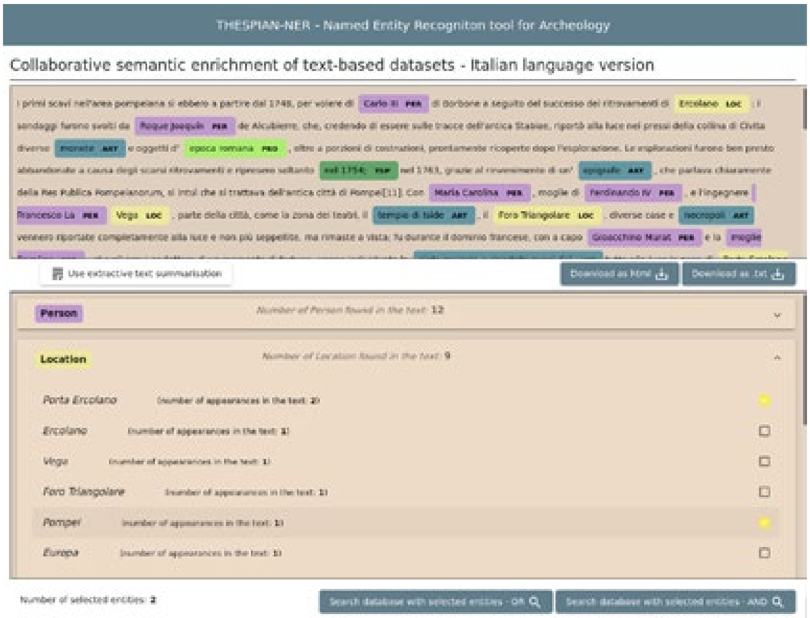
centres or groups of researchers in universities;
3. third level nodes: International partners, such as universities or research institutions from abroad.

The purpose of the platform is to create a complete ecosystem for scientific data and metadata of physical analysis on cultural heritage, modeled according to the latest ontologies and standards, and to make them interoperable with data generated by other disciplines and accessible on other platforms according to the FAIR principles for data: findability, accessibility, interoperability, and reusability [Wilkinson et al. (2016)].

The Cloud services

Firstly, a web tool dubbed Tools for HEritage Science Processing, Integration and ANalysis -Mask (THESPIAN-Mask) has been developed by the Digital Heritage Laboratory (DHLab) of CHNet, and by VAST-LAB, PIN. It consists of a web platform for assisted metadata generation and a service for persistent storage of scientific data and their metadata. The tool is based on CRMhs, an extension of the CIDOC CRM [CRM (v. 7.1.1)] ontology designed for modeling the complex entities typical of heritage science, developed by Istituto Nazionale di Fisica Nucleare and VAST-LAB PIN [Castelli et al. (2019)].

The goal of THESPIAN-Mask service is to offer a cloud environment possessing a multitude of web applications for the storage of (meta)data and their use, in light of the FAIR principles; in particular, the idea is to allow researchers to store their raw data, processed data, results, documentation and article preprints, together on the cloud, with a set of metadata based on a common, shared ontology. The system also allows data and metadata to be accessed by other researchers of the network and (re)used via a set of commonly shared web services, offered entirely on the cloud. The CHNet DHLab cloud thus constitutes a cloud for cooperative data storage and data elaboration of scientific analysis on Cultural



THESPIAN-NER - Named Entity Recogniton tool for Archeology

Collaborative semantic enrichment of text-based datasets - Italian language version

I primi scavi nell'area pompeiana si ebbero a partire dal 1748, per volere di Carlo III PER di Borbone a seguito del successo dei ritrovamenti di Ercolano LOC . I sondaggi furono svolti da Roque Joaquín PER de Alcubierre, che, credendo di essere sulle tracce dell'antica Stabiae, riportò alla luce nei pressi della collina di Civita diverse monete ART e oggetti d' epoca romana PRO , oltre a porzioni di costruzioni, prontamente ricoperte dopo l'esplorazione. Le esplorazioni furono ben presto abbandonate a causa degli scarsi ritrovamenti e ripresero soltanto nel 1754; PSP nel 1763, grazie al rinvenimento di un' epigrafe ART , che parlava chiaramente della Res Publica Pompeianorum, si intuì che si trattava dell'antica città di Pompei[11]. Con Maria Carolina PER , moglie di Ferdinando IV PER , e l'ingegnere Francesco La PER Vega LOC , parte della città, come la zona dei teatri, il tempio di Iside ART , il Foro Triangolare LOC , diverse case e necropoli ART vennero riportate completamente alla luce e non più seppellite, ma rimaste a vista: fu durante il dominio francese, con a capo Gioacchino Murat PER e la moglie

Person — Number of Person found in the text: 12

Location — Number of Location found in the text: 9

Porta Ercolano	(number of appearances in the text: 2)	●
Ercolano	(number of appearances in the text: 1)	□
Vega	(number of appearances in the text: 1)	□
Foro Triangolare	(number of appearances in the text: 1)	□
Pompei	(number of appearances in the text: 1)	●
Europa	(number of appearances in the text: 1)	□

Number of selected entities: 2

Search database with selected entities - OR
Search database with selected entities - AND

Fig. 1
THESPIAN-NER AI enhanced query service. It is possible to use automatically recognised Named Entities to build a custom query to fetch the server for similar entries.

Heritage. The use of CRMhs also makes the information fully interoperable with other CIDOC CRM compatible data and allows their integration in existing cloud environments and extended semantic graphs, such as the semantic data cloud developed by ARIADNE-plus for archaeological data.

Secondly, a web tool called THESPIAN-NER, a Natural Language Processing (NLP) tool for automatic Named Entity Recognition (NER) using deep learning techniques (namely a Convolutional Neural Network), was developed. It allows users to automatically annotate archeological documents written in Italian (either .txt or .pdf files) by identifying and labelling relevant semantic entities in the text, extract and define new metadata for the annotated documents out of them, and using them for building custom queries to fetch related records available on the CHNet database.

The Neural network was built using the python package spaCy (https://spacy.io/) and trained using transfer learning on 92 Italian written documents containing 5230 entities, annotated using the open-source program INCEpTION (https://inception-project.github.io); the 9 entities to be identified within the text have been chosen for their relevance in the archeological sector and according to their compatibility with top-level classes of CRMhs and the CIDOC CRM ecosystem. They are: Artefact, Site, Person, Timespan, Activity, Organisation, Location, Period, Biological remains.

THESPIAN-NER is available for researchers as a web tool hosted by the CHNet server in the CHNet cloud environment. Since it enables researchers to either annotate locally stored as well as remotely available documents, provided by the CHNet servers through the cloud platform, it constitutes a (micro)service aiming at fostering data (re)usability.
Finally, a tool for analysing raw data of XRF imaging is available. It allows users to fetch the XRF raw data in HDF5 format from the server, or to import their own, and retrieve both the histogram of the XRF counts and the XRF Image, shown in interactive plots that

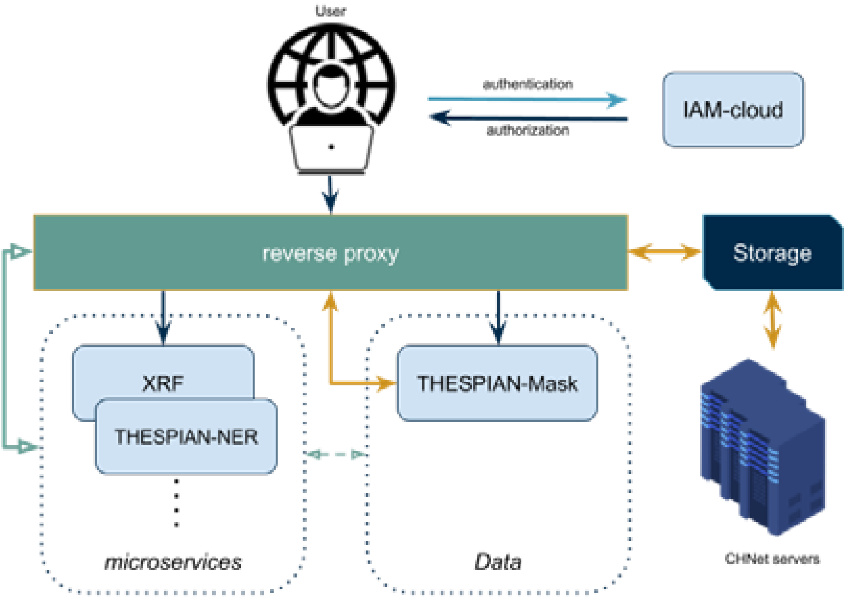


Fig. 2
Architecture of CHNet cloud

allow users to easily interact with the data representation and modify it, by adjusting all the various parameters. The raw data file content is a dictionary with key "img" and value the rank-3 tensor of shape (height, width, channel depth) representing the image.

The platforms hosting the cloud have a modular architecture based on containers: the first one hosts a web server offering a graphical web application to input and search data and institutions; the second one is a server processing requests from the web application and interacting with the database deployed within a third container. In front of these a fourth container hosts a reverse proxy that acts as an SSL/TLS terminator and enforces access control policies through a token-based authorisation mechanism relying on the IAM service [Ceccanti et al. (2019)] developed by the INDIGO project [INDIGO-DataCloud Collaboration (2017)]. The container architecture allows easy deployment of the whole cloud service. Also, the front-end part was developed using Angular with TypeScript, while the various RESTful APIs are either written using Node.js and/or Django; the latter was chosen to easily embed the deep neural network developed with Python. To persist data storage, the NoSQL database MongoDB is employed.

Conclusions

The CHNet cloud is the result of the collective efforts of PIN and CHNet: CHNet provided the hardware set-up and developed the web services; CHNet researchers are currently testing the services and populating the database with their data. PIN has contributed by developing the CRMhs ontology on which the metadata are based, by preparing the training dataset of annotated documents for THESPIAN-NER, and also by testing and populating the database; the CHNet-CNAF node has supervised the web development and has devised the cloud architecture, developing also the IAM-cloud authentication service based on OAuth2 [Hardt (2012)].

The digital infrastructure, with all its services, is currently in the pre-production stage, where only a few researchers can access the system, mainly to test the digital infrastructure and web services, ingest preliminary data and check if the design of the system fits their needs. The system will be open to all the researchers of the network for further general tests and to ingest generic scientific data on cultural heritage; the orders of magnitude the system will handle are about O(100) data report per network's node inserted in the beta testing phase, with O(10) nodes in the network, each of it consisting of O(10) researchers.

Acknowledgments

The present research has been partially funded by the European Commission within the Framework Programme Horizon 2020 with the projects ARIADNEplus (GA no. H2020-INFRAIA-01-2018-2019-823914) and EOSC-Pillar (GA no. H2020-INFRAEOSC-05-2018-2019-857650)

References

Wilkinson, M. et al. "The FAIR Guiding Principles for scientific data management and stewardship". Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

"CIDOC CRM. International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM). Version 7.1.1. http://www.cidoc-crm.org/version/version-7.1.1

Castelli, L., Felicetti, A., Proietti, F.: “Heritage Science and Cultural Heritage: a CIDOC-CRM-enabled model for Integration and Interoperability”, International Journal on Digital Libraries, Special issue on FAIR Data and Cultural Heritage data-centric research, 2019. DOI: https://doi.org/10.1007/s00799-019-00275-2

Hardt, Dick. "RFC6749 - The OAuth 2.0 Authorization Framework". Internet Engineering Task Force, (October 2012). https://tools.ietf.org/html/rfc6749

Ceccanti A., Vianello E., Caberletti M.i and Giacomini F., “Beyond X.509: token-based authentication and authorization for HEP”, EPJ Web Conf., 214 (2019) 09002, https://doi.org/10.1051/epjconf/201921409002

INDIGO-DataCloud Collaboration, INDIGO-DataCloud: A data and computing platform to facilitate seamless access to e-infrastructures, CoRR abs/1711.01981, (2017), arXiv:1711.01981

Authors

Alessandro Bombini bombini@fi.infn.it

Technological Researcher at INFN-CHNet, Florence section, is responsible for the research and development of digital infrastructure and web services for the Digital Heritage Laboratory (DHLab), within the European projects European Open Science Cloud (EOSC) - Pillar (https://www.eosc-pillar.eu/), and ARIADNEplus (https://ariadne-infrastructure.eu/).

Lisa Castelli castelli@fi.infn.it

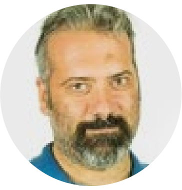
Graduated in Physics, has obtained her Ph.D. in Science and Engineering of Materials, and is a Technologist at INFN since June 2019; Lisa Castelli is an expert in the development and applications of portable X-ray-based instrumentation for the elemental characterisation of materials in cultural heritage. She is also involved in the management and organization of the activities of the INFN-CHNet network, and she actively works on the integration of the CHNet raw data, mainly in the framework of the European Projects ARIADNEplus and EOSC - Pillar.

Luca dell'Agnello luca.dellagnello@cnaf.infn.it

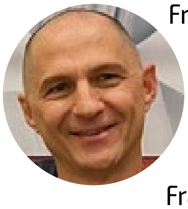
Director of Technology at CNAF, INFN. He was a Network expert for GARR network (1997-2002), Data management and Storage Group Leader at INFN-CNAF (the Italian Tier1 in LHC Computing Grid) from 2005 to 2011, and GARR Technical and Scientific Committee member from May 2009.

He is also INFN Tier1 manager since October 1st, 2011, and Director of CNAF since July 15, 2021.

Achille Felicetti achille.felicetti@pin.unifi.it



Achille Felicetti is an expert of knowledge representation and semantic web technologies and is active in the definition of ontologies and standards for Cultural Heritage. He is also responsible for the development and maintenance of digital services and infrastructures, online collaborative platforms, databases, mapping services, and web-based OS applications.



Francesco Giacomini francesco.giacomini@cnaf.infn.it
First technologist at INFN-CNAF. His activity mainly concerns the design and implementation of software components both for distributed computing systems and for physics experiments.

Franco Niccolucci franco.niccolucci@pin.unifi.it
Franco Niccolucci is a former professor of the University of Florence and the Science and Technology in the archaeology Center of the Cyprus Institute. He has been the coordinator of the projects PARTHENOS, ARIADNE, CREATIVE CH, COINS, CHIRON, and 3D-ICONS. He is the author of about 100 publications in the domain of ICT applications to CH. He was the chair of CAA2004 and founder of the VAST conference series in which he co-chaired 2000, 2003, 2004, 2011, and 2012 editions.

Francesco Taccetti francesco.taccetti@fi.infn.it
First Technologist at INFN, expert in accelerators physics and applied physics on Cultural Heritages. He is national responsible for the INFN-CHNet, the INFN network devoted to the application of nuclear techniques to the Cultural Heritages.
He is Principal Investigator for the INFN in the European projects and MIUR-financed projects, such as 4CH, ORMA, Ariadne+, MACHINA, E-RIHS, and authored more than 80 papers published in international journals.