NLP Tools for Knowledge Extraction from Italian Archaeological Free Text

Achille Felicetti
PIN, University of Florence
Prato, Italy
achille.felicetti@pin.unifi.it

Daniel Williams
School of Computing & Mathematics,
University of South Wales
Pontypridd, UK
daniel.williams@southwales.ac.uk

Ilenia Galluccio
PIN, University of Florence
Prato, Italy
ilenia.galluccio@pin.unifi.it

Douglas Tudhope
School of Computing & Mathematics,
University of South Wales
Pontypridd, UK
douglas.tudhope@southwales.ac.uk

Franco Niccolucci
PIN, University of Florence)
Prato, Italy
franco.niccolucci@gmail.com

***Abstract*:** *This paper deals with the development of advanced tools and technologies for creating relevant information and suitable metadata out of textual documentation produced by Italian archaeological research. A set of Natural Language Processing tools were developed to recognize and annotate various archaeological entities in Italian language textual reports. The CIDOC CRM is the ontology chosen for encoding resulting output, allowing for a maximum degree of standardisation of the produced metadata to guarantee interoperability with archaeological information already existing in other semantically enabled digital archives. The work took place as part of the development for the TEXTCROWD platform for the European Open Science Cloud for Research Pilot Project.*

Keywords: *NLP, NER, Italian language, archaeology, textual documents, grey literature, metadata extraction, metadata integration, standards, CIDOC CRM*

I. INTRODUCTION

European archaeological documentation consists nowadays of a multifaceted series of information, produced in different and independent ways by each of the various national and international institutions active in this discipline, by means of tools and methods that are often very different from each other. The operations of survey and excavation, and the subsequent activities of documentation and archiving, have been improved, over the years, by the use of increasingly complex and sophisticated digital tools, spanning from the graphical and geographic software used for the rendering of maps and the development of GIS, to the spreadsheets and the relational databases used for cataloguing findings and analysing the results deriving from the stratigraphic investigation.

In recent years, many European initiatives, like the ARIADNE project [1], have devoted great efforts to the integration of archaeological digital archives, mainly relying on high quality metadata generated through a series of mapping and encoding processes aimed at standardising and making them accessible and interoperable. Ontologies and terminological resources have proved to be paramount for the building of the ARIADNE Catalogue, an inventory of about 2millions archaeological datasets, a number destined to grow as new data continuously arrives. Also, the FAIR principles [2] for data structuring, on which many institutions are basing their data modelling, can be considered part of the remarkable results achieved by research in the field of openness and interoperability among cultural heritage information.

This work was funded and carried out under the framework of the EOSCpilot European initiatives (contract number: 739563).

However, many disciplines (and archaeology more than others) present a peculiar aspect in their documentation. Excavation data, for example, also comprise a huge amount of information in textual format that cannot be easily processed using traditional tools like forms, spreadsheets or relational databases. It is the rich heritage of observations and considerations that in a way constitute the very core of archaeological research, relying on the most important activity carried out by archaeologists in their research: speculation on what happened in the past by examination of what has remained in the present. Language has always been recognised by the archaeological community as one of the most powerful tools to express the richness of such speculation; thus, excavation diaries, archaeological reports and other similar documentation represent a paramount source of knowledge for scholars.

The existence of such an important information in an unstructured format, has undoubtedly represented an obstacle on the technological front, since IT tools usually require a strict and precise formatting of data in order to guarantee effectiveness in information management and retrieval.

Linguistic tools able to fill this gap by drawing "water" from the deep "well" of this knowledge and putting it in a standardised, machine understandable format, would tremendously push forward the use of digital technologies for the evolution of archaeological science. Today, powerful tools capable of "reading" a text and deciphering its linguistic structure, are fortunately available. They can work with major world languages to perform complex operations, including language detection, parts of speech recognition and tagging, syntactic and logical relationships detection. They are also able, to a certain extent and with the opportune training, to speculate in very general terms on the meaning of single words and to associate it with entities of the real world, such as people, places or events.

This paper describes a set of Natural Language Processing (NLP) tools developed to give a first answer to these issues and to test the possibilities offered by current technology for extracting knowledge from archaeological texts and creating out of it, meaningful metadata encoded in a semantic, machine readable format. Attempts at encoding Italian language textual entities in ontological format have been already carried out [3], often on corpora of Italia legal documentation [4], but to date, no effort has been devoted to the archaeological field. Thus, we have focused our investigation on Italian archaeological documents for our experiment: we have selected a corpus of Italian archaeological reports both for training and testing the

system, and chosen the CIDOC CRM ontology for codifying the resulting metadata in a semantic format, ready to be made interoperable with other semantic information created from structured data.

II. Italian Archaeology and Textual Documentation

Italian archaeology is a discipline with a history of several centuries, which has lived many lives and has gone through many phases of development ever since the 18th century, when it first became distinct from antiquarianism and from art collecting, aspiring to become an autonomous scientific discipline. Changes undergone over the years, undoubtedly affected the investigation methodology, which from time to time has always evolved. The introduction of scientific investigation methods and the benefits derived from the adoption of techniques borrowed from other disciplines, such as geology, have made it a modern discipline, while the use of new technologies for documenting excavation activity and results, has opened for it the doors of the digital world. However, textual narration has always played a central role in archaeological research, permeating the vast majority of archaeological reports from the beginning of modern archaeology until recent time.

The notorious reticence of European archaeologists to use rigid schemas for documenting their information, becomes almost systematic in Italy where, for example, the schemas and forms [5], elaborated by the Italian Ministry of Cultural Heritage to try and give an "order to the chaos", have long been neglected by archaeologists and, when used, have been used for reporting scarce or useless information in many fields, with the remarkable exception of the "notes" field, the only place deemed worthy by scholars to entrust their observations and analysis, punctually reported rigorously in a free text format.

This preference granted to the descriptive approach is, in the very peculiar case of the Italian archaeology, a merit (or, from an IT perspective, a "demerit") of the Italian rhetoric, rich in nuances and undertones and subjected, for its extreme flexibility, to be continually modelled, interpolated, enriched with new forms and expressions and with the creation of neologisms for expressing concepts suitable for specific contexts. The ease with which archaeologists, especially in the past, made use of this extraordinary tool, very often ended up transforming outstanding scientists into poets and novel writers since, for Italian scholars, the "bello stilo" (the beautiful style, according to Dante's famous words) tend to become as important as the scientific information itself. It is also worth considering that excavation diaries, notes and reports have been used over a time span of about three centuries. During such a long period, mutations occurred in the style and use of the Italian language, contextually affecting the specific and peculiar way of archaeologists to narrate their discoveries, to describe their interpretations, to make previsions and propose predictions. This phenomenon undoubtedly impacted in a very positive way on the enrichment of the Italian lexicon. On the other hand, however, when precision is required, like in the process of digital acquisition of scientific information and its interpretation, the richness of language risks to end up in an obstacle for the ambiguities and lack of "precision" that make sometimes impossible the construction of clean and clear results.

After a long period of scarce interest to these issues, the fact that in archaeology, unlike in any other empirical science, the most valuable information occurs in textual form, is nowadays widely accepted by IT community. Actually, this topic cannot be ignored if the goal is the creation of rich and valuable semantic knowledge bases of archaeological data. If what is narrated in free text format is precious information to the same extent to what is codified in databases and other structured documents, trying to extract it and make it accessible to machine in a formal way becomes essential to speed up scientific research in the archaeological field [6].

Information extraction from texts is a well-known challenge and, in recent years, many strategies have been put in place to try and overcome it. Annotation tools allowing scholars to manually mark relevant information within their reports and to map them to specific classes of modern ontologies, have been fruitfully employed in different knowledge extraction scenarios [7]; on the other side, advanced NLP tools specifically targeted at the archaeological context and able to analyse texts in an automated way, have made their appearance in the framework of various international initiatives, like ARIADNE and the PARTHENOS project [8]. During ARIADNE, a preliminary set of stand-alone NLP tools for analysing and retrieving information from archaeological reports in various languages were deployed. However, Italian was not among the languages taken into consideration at that time.

III. The EOSC framework

A big opportunity to advance NLP technology in the archaeological field was offered by the European Open Science Cloud for Research Pilot Project, EOSCpilot [9], an initiative mainly targeted to the scientific world and aimed to develop a number of high-profile pilots to show interoperability in a number of scientific domains, including archaeology. The TEXTCROWD platform [10] presented in this paper was released as one of these pilots and was intended to build an NLP cloud service capable of reading Italian excavation reports, recognising relevant archaeological entities and linking them to each other on linguistic bases. The cloud aspect of this development work is of great importance in terms of performance and interoperability: a tool running on the cloud can offer features immensely superior to the ones achievable by stand-alone applications in terms of resources management, computational performances, accessibility and reusability of results. Additionally, cloud infrastructure can be configured to set up Virtual Research Environments, a sort of virtual labs where people could work together in collaborative scenarios to address specific research questions.

The cloud version of TEXTCROWD enables researchers to generate and revise the enrichment of collections of texts availing of the automatic NLP tools provided. Additional important features concern the produced output: TEXTCROWD is actually able to generate metadata out of the knowledge extracted from the documents into a language understandable by a machine, such as RDF, to generate an actual "translation" from a (natural) language to another (artificial) one. The syntax and semantics of the latter are provided by the classes and properties of one of the main ontologies developed for the Cultural Heritage domain:

CIDOC CRM, an international standard that has become very popular and widely used in digital humanities [11].

IV. NLP AND PREVIOUS WORK

Information extraction, a major focus in NLP today, aims to extract specific elements out of a source document and make them available for further analysis [12]. Sometimes the extracted information is expressed as additional metadata for the source document and sometimes the extracted information is indicated by adding inline markup (in some form of XML) to a version of the source used for further machine processing. Named Entity Recognition (NER) is a particular subtask of information extraction that attempts to extract named entities, such as names of organisations, persons, places, from source documents. Much of the work in this area has been applied to the medical domain and web commercial applications, with limited application to archaeology. Kintigh [13] highlights the potential for the application of NLP techniques to the vast amount of archaeological reports and grey literature that is usually unavailable for meta-analysis, cross research or automated processing. Richards et al. [14] review NER work within archaeology and discuss the potential for opening access to grey literature and also more conventional publications. Their review outlines and compares the two main approaches used in NER, machine learning and rule-based. Rule-based approaches can give very accurate results but the process of setting up the rules and the terminology sources they rely on can be resource intensive. Machine learning is capable of good and efficient performance but relies on a comprehensive training set of expert annotated documents in the language of interest that encompasses all the entities to be recognised. However, this is often not available particularly in a field like archaeology where NLP approaches are a relatively recent innovation. In fact, the two approaches can be used in a complementary fashion, for example with an initial rule-based phase feeding into a machine learning process, or by using different approaches for different entities.

Most of the archaeological NER work to date has been with English language texts, although the OpenBoek project investigated the recognition of spatial and temporal entities from Dutch archaeological texts [15]. ARIADNE demonstrated the potential to broaden the application of NER within archaeology to languages other than English. Pilot NLP pipelines for English, Dutch and Swedish languages were developed, applied to archaeological grey literature reports and the outcomes analysed (ARIADNE D16.2 [16], ARIADNE D16.4 [17]). Although more work is needed for operational level application, useful results were demonstrated, and software resources provided for entities, such as artefacts, materials, dates, etc. The pipelines run on the GATE (General Architecture for Text Engineering) open source NLP platform [18]. The English language pipelines build on previous work with the grey literature library of the Archaeology Data Service [19], while the Dutch and Swedish pipelines are more exploratory (see ARIADNE D16.4 2017, for details). The ARIADNE NLP code is freely available [20].

V. TECHNOLOGY, TOOLS AND DEVELOPMENT STRATEGIES

For the purposes of TEXTCROWD, it was decided to extend the ARIADNE approach to Italian language archaeological texts. While it has been possible to draw on various general-purpose Italian language NLP processing elements, to the best of the authors' knowledge, there has not been any previous work on NER for Italian language archaeological reports. It was decided to build on the ARIADNE experience and develop a standalone tool based on the open source GATE framework [21]. The GATE framework provides reusable data structures and processing resources for creating natural language processing systems. The default pipeline provides various domain independent processing elements, with the ability to develop archaeology specific elements using rules expressed in the pattern matching language built for GATE [22]. While the lower level GATE pipeline elements are domain independent, they are not language independent and the first steps involved adapting or replacing English language modules, such as the Tokenizer, Sentence Splitter, Part of Speech Tagger and Lemmatizer with Italian language equivalents.

A. Technological Aspects of an Italian NLP Tool Development

An initial pilot NLP system for Italian archaeological reports was created within the GATE open source toolkit, taking advantage of open source Italian NLP language resources and tools. The pilot system split Italian text into sentences and tokens, performed part of speech tagging and found the stem of Italian words. The pilot system was demonstrated at a project meeting and was considered to provide a suitable basis for subsequent development towards a full Italian archaeology NER system.

An initial candidate set of Italian archaeological entities (for the demonstrator NER work) had been identified in discussions between the Italian archaeology team and the NLP developers. This was reassessed in the first evaluation exercise and refined. Some entities were omitted that currently did not appear to have sufficient terminology (vocabulary) resources associated with them. Some entities proved to be ambiguous and were defined more precisely (for example, site vs monument, object vs material). This proves to be a complex issue (see discussion in section V*C*). The final set of elements for the NER system comprised Artefact/Monument, Colour, Material, Period, Place, Person, Site, Technique and Timespan entities to be identified in Italian archaeological text.

Expert annotation by Italian archaeologists of a sample of archaeological reports, guided by instructions for annotators produced for the project, in order to feed into and help evaluate the NLP software tools produced for the demonstrator, continued throughout the development effort. Involving human annotators is always a very time-consuming exercise and resources need to be budgeted for in any project of this kind. The procedure to use the human annotations and ingest them into the NLP system was facilitated by the fact that the expert archaeologist annotators were able to install and use the GATE framework for the annotation work.

An initial set of Italian language resources to use (gazetteers, glossaries, thesauri, etc) had been identified by the Italian archaeological team (see below, sections V*B* and V*C*). The vocabulary coverage was considered as part of the first evaluation exercise, where it became clear that further enhancement was required and, in some cases, pre-

processing of the original vocabulary sources. Again, we return to this in the concluding discussion.

The default GATE framework comes with a pipeline of low level English language processing elements. To build a named entity recognition system for another language, low level processing of that language needs to be added. In some cases, this has already been produced for that language and NLP framework. In our case, this had to be added by making use of available Italian NLP resources. Low level processing includes tokenizing the document with annotations that span each word (this includes detecting abbreviations and use of special characters in some word forms) and splitting a text into discrete sentences. An annotation is a data structure provided by the GATE framework that acts as a form of metadata about the document. Special sections of a document, such as tables and bibliographic references, can be problematic for tokenizing. In order to handle singular, plural and other syntactical forms of a word, it is helpful to find the stem value (without the word ending) of each word in the document and to use the stem value when comparing potential lookups of words in the document with the glossaries for the different entities. This is useful for entities such as artefacts but less useful for place name and period entities, where typically stemming is not applied. Another important step is identifying the relevant part of speech and adding the POS tag (noun, pronoun, verb, etc) to each "token" annotation for each word in the document. Knowing the part of speech allows the development of higher level syntactical rules that begin to address word phrases and particular patterns found in the application domain (we see below). Typical POS taggers usually reply on statistical machine learning techniques applied over large document collections or corpora. The automatic POS tagging is not necessarily always correct particularly where an extensive training corpus has not been available.

The GATE framework can parse and extract human readable content from various documents formats including PDF and HTML files. Therefore, the archaeological text PDF documents are pre-processed by GATE before being passed on to the NER system. Some additional pre-processing was added to deal with hyphenated words and footnotes.

The low-level processing algorithms for the initial pilot employed a set of OpenNLP (https://opennlp.apache.org/) components for the Italian language. Following the initial evaluation, it was decided to adopt some of the low level Italian language processing resources produced for the (EU funded) OpeNER (http://www.opener-project.eu/) project, such as language identification, tokenization and part of speech tagging. Both sets of 3rd party NLP resources were combined in the final NER system. In addition, the TEXTCROWD NER pipeline is able to directly annotate "Place" and "Person" entities using the OpeNER project's webservices.

In addition, a Gazetteer was produced to check if a word (stem in most cases) is present in any named entity glossary, together with more complex rules developed specifically for the Italian archaeology domain. These rules use information such as the part of speech tag of a Token and whether a Token is present in named entity list to identify named entities.

Figure 1 shows the Artefact (and Monument), Period, Place, Site and Timespan annotations produced by executing the TEXTCROWD NER system on a pdf archaeological report using the GATE software, the entities are coloured are green, purple, cyan, pink and red respectively.

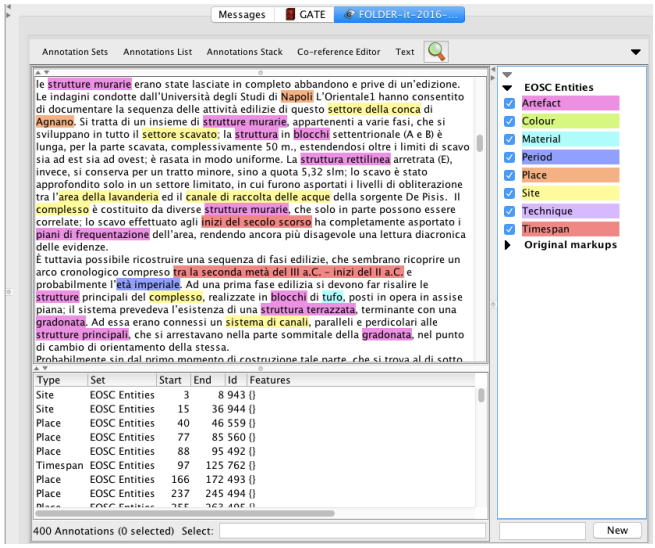

Figure 1: Sample output from NER pipeline

Some effort was spent on the development of specialised rules for artefact, timespan and period entities that were based on common patterns observed in the expert archaeologist annotations or from directly supplied examples (for timespan and period expressions). The artefact rules addressed multi-word phrases that make use of syntactical patterns and artefact vocabulary lookups. The patterns are empirically based on an analysis and selection of the most common patterns involving artefact vocabulary from the expert annotations.

B. The Creation of a Manual Annotated Archaeological Corpus

Exploiting the potentiality of rule-based and machine learning combined approach, we had the possibility to evaluate previous pipelines designed for archaeological documentation and define an efficient work strategy (ARIADNE D16.4 [17]). The creation of an archaeological corpus has been a crucial and preliminary operation to produce a first dataset of grey literature reports for rule-based approach.

The Fasti Online FOLD&R Italy Series [23] has been chosen as archaeological online database, representing one of the best providers for preliminary and final reports on excavations from 2000 onwards. FASTI documents are well structured and clean, they proved to be the most suitable kind of documents for our purposes.

Reports have been selected using different criteria and ensuring a variety of provenance, chronology etc. Five variables have been defined to support the selection within a huge amount of reports. This variety could ensure the possibility to find different linguistic patterns and a variegated lexical assortment. In details, the variables are: Place (excavation reports conducted in different locations); Period (excavation reports conducted in different historic periods); Type of Institution (excavation reports conducted by different actors); Archaeological Process Phase (reports

produced in different phases of archaeological excavation); Year (attributable to the publication year).

The Corpus contains 30 reports, composed of an average of ten pages, covering almost the whole national territory, the chronological coverage starts from Bronze Age to Post Renaissance. Excavations and related documentation have been produced from different archaeological investigators, i.e. private companies, universities, authorities ("Soprintendenza" in Italian), during different phases, i.e. preliminary field survey, final excavation, desk-based assessment.

The subsequent step was the manual annotation using GATE application and following the "Instructions for Annotators" provided during the ARIADNE project (ARIADNE D16.2 [16], ARIADNE D16.4 [17]).

This operation has been achieved by a team of three archaeologists, in order to ensure a different approach during the evaluation of ambiguous lexical attributions (i.e. the ambiguity in the meaning of some Italian archaeological concepts, like Artefacts, Monuments and Sites). From a semantic point of view, the process of standardisation has been a complex operation. Annotation means isolating and marking keywords within each text by assigning them to the eight categories (see V first paragraph), by preserving their original archaeological meaning. So far, in the ARIADNE project, the annotation has been conducted only for Germanic languages; the application of rules and instructions to a Romance language, like Italian, needs to be performed in a most flexible and elastic way, considering the typical rhetorical nuance characterizing these languages, particularly in expressing time span patterns.

C. Thesauri and Terminological Resources

Identifying existing terminological resources, combined with the production of new lists of terms, has been a parallel activity. Thesauri, in fact, were useful not only for the archaeological team in order to have a general linguistic pipeline to follow during the annotation phase, but mostly to allow the enrichment of them when they were lacking.

To provide Gold Standards for Italian archaeological grey literature and ensure a substantial amount of official vocabularies, we conducted survey within the ICCD (the Italian Central Institute for Cataloguing and Documentation) [5]. Among the other resources provided by ICCD, the RA Thesaurus (Archaeological Findings Thesaurus) proved to be one of the most useful and strategic ones for the description of archaeological artefact, due to its completeness.

During this phase, the most remarkable difficulty has been defining a semantic clear separation between Artefact and Site concepts, because their meanings often overlap. We decided to split them into two separate lists, considering the wider concept of Artefact as a Man-Made Object (following the CIDOC CRM approach), including Monuments and Buildings. The Site list has been significantly reduced, by leaving within it only the terms that indicate extended archaeological areas, such as "Terrazzamento a scopo di consolidamento" (Terracing for consolidation purposes).

A general overview of all the linguistic resources we used for each category is provided below:

Artefact:

1. ICCD RA Thesaurus.: thanks to its hierarchical schema, RA Thesaurus contains information about Archaeological Objects, Functions, Morphologies and Parts. We decided to consider the first three cathegories as a unique linguistic entry within a textual document, while each Part is considered as an Artefact itself. For example: "Cintura/per la sospensione delle armi/multipla" (Belt/for bearing arms/multiple), consists of Object+Function+Morphology. Function and Morphology cannot express a stand-alone meaning. Each entry within the Part category, on the contrary, represents an object itself, for example "Borchia" (Stud).

2. ICCD NU- Object Thesaurus: this terminological list provides numismatic terms.

3. ICCD List of Monuments, the document contains a set of building archaeology terms (i.e. "Ponte di Diocleziano" – Diocletian's bridge).

4. ICCD A - Architectural and landscape heritage, where it is important to integrate the architectural heritage to the Artefact list.

5. ICCD SI – Site, this vocabulary has been processed and divided into an artefact list and site list, as specified above.

Colour: Wikipedia Italian List of Colours [24].

Material: ICCD OA – Art Objects – Material and Technique. This vocabulary, designed for art objects, has been adapted to archaeological materials.

Period:

1. PICO Thesaurus, provided by CulturaItalia [25], has been chosen to represent a categorized standardization of historical periods.

2. PeriodO [26], a gazetteer of period definitions for linking and visualizing data.

Place:

1. Geonames [27], a geographical database, which covers all countries and contains over eleven million names, for actual places.

2. Pleaides [28], a community-built gazetteer of ancient places.

Site:

1. ICCD SI – Site, including only the split site list.

2. ICCD List of Sites, containing archaeological areas and parks, such as "Area Archeologica di Paestum" (Paestum Archaeological Area).

Technique: ICCD OA – Art Objects – Material and Technique. This vocabulary, designed for art objects, has been adapted to archaeological techniques.

Timespan: for this category we have built an ad hoc list of recurring patterns that have been deduced from the analysis of the annotated documents. For example: the expression "Tra la seconda metà del III secolo a.C. e l'inizio del II sec. a.C." (Between second half of the III Century B.C. and the start of the Second Century B.C.), even if presented in a narrative way, is full of salient information that the tool

would not have detected without explicit identification of this kind of patterns.

All these vocabularies have been integrated with "Annotations to Gazetteers" lists, designed for each category, containing all the terms extracted from the manual annotations, which were fundamental for identifying those generic terms that often occur in the common language of the reports, but which are not included in standard vocabularies. For example: "Ambiente circolare" (circular room).

D. D4Science and the Cloud

The TEXTCROWD service, made available via the D4Science infrastructure [29], allows users to upload and store textual documents in a personal cloud folder, perform NLP and NER operations, trigger the semantic enrichment process and obtain the output CIDOC CRM information in RDF. Results can be uploaded in a triple store or in another semantic enabled system and reused within the same context or in another VRE (Virtual Research Environment) scenario on the same cloud.

The NER system for Italian language archaeological text developed for TEXTCROWD annotates Artefact, Colour, Material, Period, Place, Person, Site, Technique and Timespan entities. After executing on Italian archaeological text, the algorithm will produce CIDOC CRM RDF, GATE XML, Inline HTML and a list of annotations.

The TEXTCROWD NER tool's algorithms can be used via D4Science's website or used as a webservice. To access the NER algorithms, you must have a user account on D4Science's website and be able to access the "EOSC Pilot" virtual research environment. The "String" and "File" algorithms are the same but accept different kinds of user input. The "String" algorithm accepts plain text from the user (either typed or text copied into the text field) and the "File" algorithm accepts a pdf file.

An example image of an "inlineHTML" file, showing the various fragments of texts recognised in a report by the tool and the associated conceptual entities in different colours is presented below.

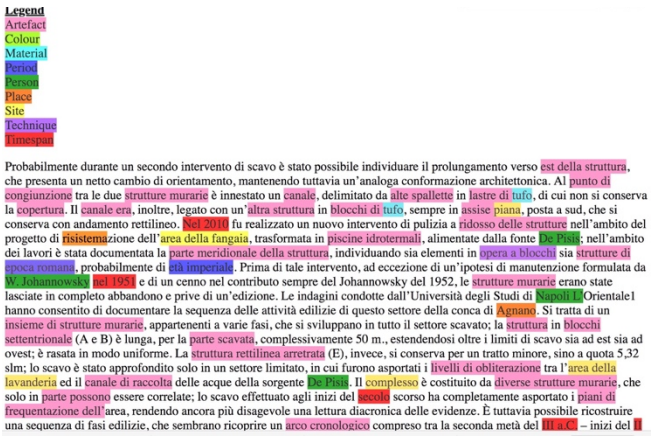

Fig 2: Annotated text in HTML format

The NER pipeline was made executable via the command line as a GATE application and bundled with other dependencies such as GATE's libraries, so that it could be deployed on the D4Science's infrastructure. D4Science offers a modular data infrastructure service, operated and maintained by CNR-ISTI, which is based on Virtual Research Environments created by each user that contain cloud storage, data catalogues and analysis algorithms. It is built on the open source gCube software system and supports services based either on a web-based GUI (Graphical User Interface) or programmatic access via an API (Application Programming Interface).

The NER webservices that were created for the EOSC pilot (April 2017 – March 2018) were first created in a prototype VRE as they were being developed and then transferred over to the 'TextCrowd' VRE by D4Science when ready for use. Initially D4Science functionality focused on the statistical R programming language but during the project additional functionality was released so that other programs could run on the cloud platform, which proved useful. DataMiner is a section of the D4Science website that allows users to use algorithms, see Fig 3 below.

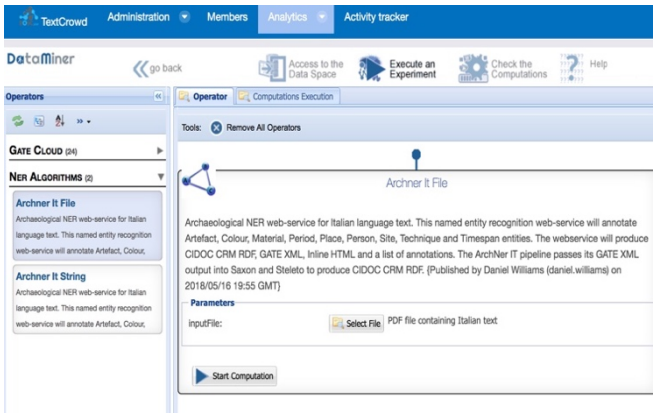

Fig 3 D4Science environment for TEXTCROWD

The CIDOC CRM based RDF output of the NER webservices need the Saxon (http://saxon.sourceforge.net/) and STELETO [30] software to transform the XML file immediately from GATE and produce the resulting CIDOC CRM RDF output file.

The web service basis of the D4Science platform allowed the NLP application to augment the GATE pipeline with the specific Italian NLP resources of OpeNER project. For the pilot project, this was achieved by external web service calls from within D4Science.

VI. TEXTCROWD TESTS AND RESULTS

Various tests have been carried out at the end of the development process, by processing archaeological reports of different topic, length and structure. Results have demonstrated a satisfactory capability of the tool to identify most of the relevant archaeological entities present in the text and convert them to suitable metadata for the semantic description of the annotated documents. TEXTCROWD has proved to be very clever in identifying actors and institutions, both ancient (e.g., names of people or tribes, emperors, warriors, builders and so forth) and modern (e.g.: archaeologists, universities, superintendences and other archaeological institutions) ones. A particular aptitude has been observed in the recognition of temporal entities and

periods, even the ones reported in the articulated and often cautious expressions usually loved by archaeologists. Dates like “ultimo quarto del V sec. a.C.” (= 425-400 BC) or “primo decennio del II sec d.C.” (= 100-110 AD), are easily identified by the tool and correctly assigned to the corresponding time spans. There is obviously no problem in retrieving periods having a corresponding entry in the thesauri used for training the tool (PeriodO and ICCD *in primis*), even in presence of variations in terms like “Periodo Romano”, corresponding to “Epoca Romana” in thesauri.

The use of gazetteer and similar lists of ancient and modern places, combined with the NER framework on which TEXTCROWD relies, makes the identification of places and place names very straightforward. Ancient cities, disappeared, no longer inhabited or currently existing with different names (and thus, having no specific entries in modern gazetteers), have been discovered by the tool with a very good degree of approximation. The tool is however still unable to guess co-references between place names like for instance “Pythecoussai” (ancient name) and “Ischia” (modern name), both referring to the same island. Thanks to NER mechanism, good performances have been demonstrated by the tool also in identifying popular Italian archaeological sites. Some difficulties in recognizing minor or new archaeological sites, usually “invented” and assigned on the flight when specific places are designated for excavation (e.g. “Villa Romana di Settefinestre”), looks perfectly normal given the high degree of variability of these names. Such small difficulties could be overcome by training the tool with a sufficient number of examples to improve its sensibility towards these entities.

As it is quite easy to guess, artefacts and their features (i.e.: materials, shapes, colours, production techniques) are actually the most problematic group of entities to manage, mainly because of the huge variety of names, definitions and derived expressions for their description created from time to time by archaeologists for their identification. There is often some ambiguity in natural language as to whether a word refers to a material per se or an (unnamed) object made of that material. The ICCD RA thesaurus provides thousands of concepts but, despite this abundance of names, Italian archaeologists actually seem to give their best in imagining artefact types and inventing expressions to describe objects. TEXTCROWD does its best to try and identify as many entities of this kind as possible; but names like “piede troncoconico di coppa corinzia” (conic-shaped foot of Corinthian cup) are not recognized. The pattern-based artefact rules have made some progress. However, they are not always consistent, being dependent on accurate results from the lower level NLP components, POS tagging and word variations. For example, phrases such as “frammenti anforacei” (fragment of various amphorae), or “mattoni sesquipedali”, (bricks measuring one *sesquipeda*, an ancient measurement unit equal to 1 and half Roman feet) are sometimes recognized and sometimes not. On the other hand, the reasonably complex phrase "frammento di anfora samia" (fragment of a amphora coming from Samos island) is recognized. Thus, we believe that further training with a reasonable number of annotated texts, combined with a more elaborated ruleset (possibly feeding into a machine learning component), would provide TEXTCROWD with the capability of mirroring the patterns used by archaeologists when minting expressions for such entities in order to identify them.

VII. Conclusions and Future Work

From the point of view of the encoding of resulting information, the TEXTCROWD framework has demonstrated a high capability to create rich and useful metadata out of the entities identified within the text. This can support enhanced cross search across different language reports and datasets, including grey literature, facilitating meta research and comparative studies.

The CIDOC CRM triples generated by the tool are semantically coherent and clean results, thanks to the perfect adherence to the classes provided by the ontology for describing the various entities involved in the narrative flow of each document. Encoding textual entities by means of CIDOC CRM classes, makes archaeological information generated by TEXTCROWD FAIR-enabled and immediately ready to be shared with (and thus, to enrich) other RDF open data based on the same semantic model and available through public archaeological services, like the ARIADNE Registry.

CIDOC CRM is an ontology based on events, a set of entities very difficult to identify within a text, but of paramount importance for linking together all other entities (actors, places, objects and so on) in order to build “semantic stories” and to define advance layers of semantic information. At its current stage of development, TEXTCROWD is unable to establish such correlations via events, and thus, to exploit the full potential of the ontology. The pattern-based rules begin to point in this direction. Tests on such functionality are already planned for the future.

We plan to conduct a systematic analysis and evaluation of the performance to date. Other future work includes the refinement of the pattern rules to take account of identified common variant expressions in Italian archaeological writing, including statements of negation - eg "no evidence of early Roman activity was found".

One final development was the incorporation of vocabulary used in the expert annotated training corpus that was not represented in the domain vocabularies and thesauri used as a starting point. This increases the vocabulary coverage to include more of the terminology employed in the archaeological reports and also serves as a resource to enrich the coverage of the corresponding thesauri.

TEXTCROWD has shown itself to be extremely useful as a demonstrator of the importance of EOSC for scientific research in the heritage domain and is ready to be made available to the broader scientific community. Porting the framework to English, as well as to other languages for which archaeological vocabularies and appropriate resources are available, will be straightforward and, with some adaptations, TEXTCROWD is ready to be adapted to other, completely different domains where appropriate vocabularies and ontological tools are available.

Acknowledgment

Thanks are due to Ceri Binding (University of South Wales), who assisted in the development of timespan rules and STELETO data conversion. We are also grateful to Anna Reccia and Francesca Forte (PIN, University of Florence), for their priceless effort and unparalleled accuracy in annotating the archaeological reports used for training the tool.

REFERENCES

[1] ARIADNE Project: http://www.ariadne-infrastructure.eu

[2] M.Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship" . Scientific Data, 3, 160018, 2016.

[3] F. Dell'Orletta, A. Lenci, S. Marchi, S. Montemagni, V. Pirrelli, and G. Venturi, "Dal testo alla conoscenza e ritorno: estrazione terminologica e annotazione semantica di basi documentali di dominio." AIDA Informazioni, Proceedings of the National Conference Ass.I.Term I- TerAnDo, AIDA, n. 1-2/20085:185–206, 2008.

[4] A. Lenci, S. Montemagni, V. Pirrelli, and G. Venturi, "Ontology learning from italian legal texts." In Proceedings of the 2009 Conference on Law, Ontologies and the Semantic Web, pages 75–94, Amsterdam, The Netherlands. IOS Press, 2009

[5] ICCD Standards: http://www.iccd.beniculturali.it/index.php?it/473/standard-catalografici

[6] F. Niccolucci et al., "Managing Full-Text Excavation Data with Semantic Tools", VAST 2009. The 10th International Symposium on Virtual Reality, Archaeology and Cultural Heritage, 2009.

[7] A. Felicetti, Teaching Archaeology to Machines "Extracting Semantic Knowledge from Free Text Excavation Reports", ERCIM News 111, October 2017.

[8] PARTHENOS Project: http://parthenos-project.eu

[9] EOSCpilot: http://eoscpilot.eu

[10] TEXTCROWD: https://eoscpilot.eu/science-demos/textcrowd

[11] CIDOC CRM: http://cidoc-crm.org

[12] J. Cowie and W. Lehnert, "Information extraction", Communications ACM, 39(1), pp. 80–9, 1996

[13] K. Kintigh,"Extracting Information from Archaeological Texts. Open Archaeology"; 1: 96–10, 2015

[14] J. Richards, D. Tudhope and A. Vlachidis, "Text Mining in Archaeology :Extracting Information from Archaeological Reports." Mathematics in Archaeology (eds. Juan Barcelo; Igor Bogdanovic). Florida. p. 240-254 12. DOI: 10.1201/b18530-15, 2015

[15] H. Paijmans, S. Wubben, "Preparing archaeological reports for intelligent retrieval, in Layers of Perception." Proceedings of the 35th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA) Berlin, Germany, April 2-6, 2007, (ed. A. Posluschny, K. Lambers and I. Herzog), 212-217, Kolloquien zur Vor- und Frühgeschichte Band 10, Bonn, 2008

[16] ARIADNE D16.2., "First Report on Natural Language Processing.2 http://www.ariadne-infrastructure.eu/Resources/D16.2-First-Report-on-Natural-Language-Processing, 2015

[17] ARIADNE D16.4.,"Final Report on Natural Language Processing." http://www.ariadne-infrastructure.eu/Resources/D16.4-Final-Report-on-Natural-Language-Processing, 2017

[18] H. Cunningham, D. Maynard, K. Bontcheva. and V. Tablan, "GATE: A framework and graphical development environment for robust NLP tools and applications", Proc. 40th Annual Meeting of Association for Computational Linguistics; New Brunswick, pp 168-175, 2002

[19] A. Vlachidis, D. Tudhope D, "A knowledge-based approach to Information Extraction for semantic interoperability in the archaeology domain." Journal of the Association for Information Science and Technology, 67 (5), 1138–1152, Wiley, 2016

[20] ARIADNE NLP., "English, Dutch, Swedish rule-based Natural Language Processing pipelines." http://ariadne2.isti.cnr.it/index.php/services, 2016

[21] GATE. GATE NLP Framework. https://gate.ac.uk

[22] H. Cunningham, D. Maynard, and V. Tablan, "JAPE a Java Annotation Patterns Engine (Second Edition)" [online] Technical report CS--00--10, University of Sheffield, Department of Computer Science. Available at http://www.dcs.shef.ac.uk/intranet/research/resmes/CS0010.pdf, 2000

[23] FastiOnline FOLD&R: http://www.fastionline.org/folder.php?view=home

[24] https://it.wikipedia.org/wiki/Lista_dei_colori

[25] http://www.culturaitalia.it/

[26] PeriodO: http://perio.do/

[27] Geonames: http://www.geonames.org/

[28] Pleiades: https://pleiades.stoa.org/

[29] D4Science: https://www.d4science.org/

[30] STELETO. "STELETO data conversion tool." http://ariadne2.isti.cnr.it/index.php/services, 2016

All web references were checked on 01/06/2018