The Archaeo-Term Project: Multilingual Terminology in Archaeology

Giulia Speranza, Raffaele Manna, Maria Pia di Buono, Johanna Monti
UniOr NLP Research Group
"L'Orientale" University of Naples
Italy
{gsperanza, rmanna, mpdibuono, jmonti}@unior.it

Abstract

In this paper, we present the Archaeo-Term Project, along with one of its first efforts in enhancing multilingual access to Archaeological data, making available a resource of Archaeological terms within the framework of YourTerm CULT project. In order to enhance and promote the use of a terminological common ground across different languages the Archaeo-Term multilingual Glossary is intended both for scholars, experts in the field, translators and the general public. Its first release contains terms in Italian, English, German, Spanish and Dutch together with PoS, definitions and other linguistic information. This paper presents the data and the methodology adopted to create the glossary as well as the evaluation of the first results.

1 Introduction

Languages for Special Purposes (LSP) have their roots in the need of communicating specialised and technical knowledge within a restricted group of domain experts.

From a linguistic perspective, LSP are mainly characterised by the use of specialised terminology, which is usually monosemous for the principle of clearly defining concepts and avoiding miscommunication and can often result opaque and unintelligible to laypeople (Gotti, 2008; Cabré, 1999; Faber and Rodríguez, 2012; Crystal, 1997). In fact, for these reasons, it is often necessary to modulate specialised languages when both oral and written communication takes place between expert and non-experts, in order to ease the didactic and informative functions of communication (Cortelazzo, 1994).

The language used in the domain of Cultural Heritage (CH), and its sub-domains, such as Archaeology, shares many points with other LSPs, such as the presence of technical terminology, terms of Greek and Latin origins, re-semantisation of common words into specialised domains of knowledge, complex multiword expressions, to mention a few. Nonetheless, it has been traditionally less investigated if compared to, for example, the language of medicine or law, which are considered soft disciplines too. As a consequence, except for a few felicitous examples (see Section 2), language resources and especially terminological resources, in this domain, are still needed.

Language resources such as glossaries, thesauri, dictionaries and term-banks are invaluable sources for language experts, translators, learners, among others. Their development can often be demanding and time-consuming, especially when carried out manually.

Specialised domain resources are even more challenging because their creation also needs the validation of experts in the domain of knowledge.

In this paper we present our work aimed at the creation of a multilingual glossary of archaeological terms, which is useful in many application scenarios from Machine Translation (MT) to Natural Language Processing (NLP).

The remainder of the paper is organized as follows: Section 2 describes related work and, following this, Section 3 presents the Archaeo-Term Project's aims and the creation of the multilingual glossary of archaeological terms, along with the description of the starting data used so far, namely the ICCD Thesaurus, and the methodology applied to extract multilingual data from the Getty AAT. To complete this section, we illustrate the first results together with their evaluation. Finally, the paper ends with the conclusions and the future work.

2 Related Work

Terminology, as several scholars pointed out (Wright et al., 2010; Melby, 2012), may sometimes result in a heterogeneous activity involving different formats, data models and practices; therefore, in order to support the sharing and the reuse of terminological resources, several standard formats have been developed, such as TermBase eXchange (TBX) (Melby, 2015).

More recently, with the spreading of the Semantic Web Technologies, many language resources are being released in compliance with the Linked Open Data (LOD) principles, using formalisms such as SKOS and Ontolex-Lemon, which are based on the Resource Description Framework (RDF), for representing glossaries, vocabularies and taxonomies (Chiarcos et al., 2013).

In the field of CH some language resources have been released during the years, both monolingual and multilingual. Among the multilingual resources, the most referred one in this domain is the Art & Architecture Thesaurus (AAT)[2], developed and maintained by The Getty Research Institute. It is a multilingual thesaurus used to describe art, architecture, decorative arts, material culture, and archival materials, which can be accessed through a web interface or via its LOD version (JSON, RDF, N3/Turtle, N-Triples), as well as XML and relational tables.

Another multilingual terminological project on CH is the iDAI.vocab[3], a controlled vocabulary specifically designed for archaeological terms available in several languages, developed by the German Archaeological Institute (DAI).

Many other glossaries and thesauri have been created as monolingual resources for cataloguing purposes. Such as the vocabularies developed by the FISH (Forum on Information Standards in Heritage)[4] and maintained as LOD resources by the Heritage Data[5] for English, or the thesauri and controlled vocabularies developed by the Italian Institute for Cataloguing (Istituto Centrale per Il Catalogo e La Documentazione - ICCD) [6].

The ICCD has also started, in 2017, the ArCo project[7] together with l'Istituto di Scienze e Tecnologie della Cognizione (ISTC) del CNR, in order to make available data from the General Catalogue of Cultural Heritage according to the LOD principles (Carriero et al., 2019b; Carriero et al., 2019a).

Some glossaries are also released by the museums or cultural institutions such as the British Museum's Object Names Thesaurus[8].

In the field of Cultural Heritage in general, and particularly, in archaeology, it is worth mentioning the ARIADNE Project (Meghini et al., 2017) which provides a portal for the collection of data and resources in order to overcome the fragmentation of archaeological data repositories of all types.

3 Archaeo-Term project

The Archaeo-Term project of the UNIOR NLP Research Group[9] of the University of Naples "L'Orientale" is part of the YourTerm CULT initiative[10] in partnership with the Terminology Without Borders program fostered by the Terminology Coordination Unit (TermCoord)[11] of the European Parliament's Directorate-General for Translation (DG TRAD). Among the different projects, YourTerm CULT is specifically designed to operate in all aspects of culture.

The Archaeo-Term project has been launched to fill the gap in an important field which takes us back to the roots of European culture and history, namely Archaeology.

The project aims at improving the accessibility of the archaeological information available in various sources (scientific papers, texts addressed to general audiences, web sites, structured databases, etc.) by creating language resources useful to NLP and MT tasks across languages. This will ease the availability of the information that can be used to structure and connect different types of knowledge bases together, both structured databases and un-

[2] https://www.getty.edu/research/tools/vocabularies/aat/about.html

[3] https://archwort.dainst.org/it/vocab/index.php

[4] http://www.heritage-standards.org.uk/terminology/

[5] https://www.heritagedata.org/blog/vocabularies-provided/

[6] http://www.iccd.beniculturali.it/it/strumenti-terminologici

[7] http://stlab.istc.cnr.it/stlab/project/arco/

[8] http://terminology.collectionstrust.org.uk/British-Museum-objects/

[9] https://sites.google.com/view/unior-nlp-research-group

[10] https://yourterm.org/yourterm-cult/

[11] https://termcoord.eu/

structured text collections.

Indeed, although some scientific communities felt the need to structure their knowledge by means of thesauri or ontologies, the scenario is still very fragmented as posed by Felicetti et al. (2018). Nowadays, European archaeological documentation consists 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. Thus, there is still the need to establish a terminological common core shared across languages.

In this scenario, the Archaeo-Term project tries to contribute to the improvement of scientific cooperation and advancements by attracting both academia and museums from different countries in the creation of a wide multilingual terminological resource in Archaeology. With this aim in mind, one of the first results of this project is a multilingual Glossary of archaeological terms which is mainly useful for the multilingual digitalisation efforts of the museums, but also to scholars, translators and the general public.

3.1 Data and Methodology

For the creation of the Archaeo-Term multilingual glossary, we start from the RDF/SKOS version of the Italian ICCD Thesaurus[12], one of the best practices adopted by the Italian Ministry of Cultural Heritage (MiBAC) to publish institutional information as LOD, in order to be easily findable, reused and freely shared. It contains 1,059 Italian terms which are linked to the LOD version of the Getty AAT[13], by means of the `skos:closeMatch` property pointing to the Getty URIs (Figure 1). This property is used to link two similar concepts that can be used interchangeably in some information retrieval applications (Cfr. SKOS Recommendation 18 August 2009). We choose to extract the information stored into the Getty AAT because it is a valuable and trustworthy resource, created by experts in the field.

The exploitation of the ICCD resource to read URIs pointing to Getty AAT contributes to build our multilingual glossary of archaeological terms along with the corresponding definitions and sources in other languages, namely English, Spanish, German and Dutch. Among the many languages available in the Getty AAT, we decide to use for our glossary those mentioned above since they show the best coverage in terms of linguistic equivalence (translations) starting from the Italian terms in the ICCD thesaurus.

In order to perform this, we use the Getty AAT SPARQL Endpoint[14] to access term related information by means of setting queries. In detail, the querying process consists of a matching operation between the results of integrated queries in the AAT SPARQL Endpoint.

We first use a query capable of parsing the ICCD resource and reading each URI which refers to the corresponding English archaeological term. In fact, in the Getty AAT, English terms and other available corresponding terms in different languages are represented as equivalent terms by means of the `skos:prefLabel` property[15] and as alternative terms in `skos:altLabel` property[16]. Both properties carry one lexical value and one language tag, associated with the lexical value, for each URI.

Since we try to extract corresponding terms in different languages, we then perform a further query able to extract archaeological equivalent terms along with their language tags and alternative terms along with language tags for each available language per URIs.

In addition to this, we set another query able to read URIs and collect corresponding definitions and sources along with their language tags, (both contained in the `skos:scopeNote` property)[17]. As a first result of such a query looping over ICCD URIs, we collect archaeological terms, definitions and sources. These queries guarantee the exploitation of the Getty AAT resource but, regardless of the language tags, also a combination of each term value associated with each definition and source value (present in the `skos:scopeNote`).

[12] `https://github.com/ICCD-MiBACT/Standard\-catalografici/blob/master/strumenti-terminologici/beni\%20archeologici/ICCD\_Thesaurus\_definizione\%20del\%20bene\_reperti\%20archeologici.rdf`

[13] For the mapping process see the ARIADNE project described in Felicetti et al. (2015)

[14] `http://vocab.getty.edu/sparql`

[15] `https://www.w3.org/2012/09/odrl/semantic/draft/doco/skos_prefLabel.html`

[16] `https://www.w3.org/2012/09/odrl/semantic/draft/doco/skos_altLabel.html`

[17] `https://www.w3.org/2012/09/odrl/semantic/draft/doco/skos\_scopeNote.html`

To the best of our knowledge, in the AAT we did not find a direct link between the different language terms values (stored in skos:prefLabel and skos:altLabel and the different language literal values (definitions and sources in skos:scopeNote) represented for the same URI. Therefore, to build our multilingual glossary we rely on a matching operation between URIs and language tags related to term values (represented in skos:prefLabel and skos:altLabel), definitions and sources (both represented in skos:scopeNote).

In particular, starting from a combination of all term values and literal values (definitions and sources) per language present for an URI, we apply a matching operation able to select only the terms, definitions and sources concerning the same language based on the reference URI. This matching operation allows us to recognise and organise archaeological terms and their literal values, that is definitions and sources, pertaining to the same language for each archaeological term identified by URI.

3.2 Results and Evaluation

Once the queries steps are performed, we first replace retrieved URIs with numeric IDs in order to provide an identification code for each entry of our glossary; then we build monolingual tables for each language mentioned above and a multilingual synoptic table.

For monolingual tables, we automatically classify in separated tables all retrieved data based on the language tag for each term entry. On the other hand, we align the terms in the different languages based on the shared ID to build the multilingual synoptic table.

In detail, the Glossary first release[18] is organised as follows:

- For each language forseen in the glossary (Italian, English, Spanish, German and Dutch) there is a dedicated monolingual table, named after the corresponding language locale (e.g., IT for Italian, EN for English) which contains 8 fields (ID, Singular Term, Plural Term, Qualifier[19], PoS, Alternative Terms, Alternative Terms Qualifier, Definition and Source) as shown in figure 2.
- a multilingual synoptic table contains all the languages singular terms, which are linked to one another by means of the IDs. This multilingual table aims at providing a comprehensive overview on the equivalent terms across the languages.

During the evaluation phase, we noticed that 9 Italian terms had two equivalent English terms in the Getty AAT, marked by two closeMatch URIs to the AAT instead of just one.

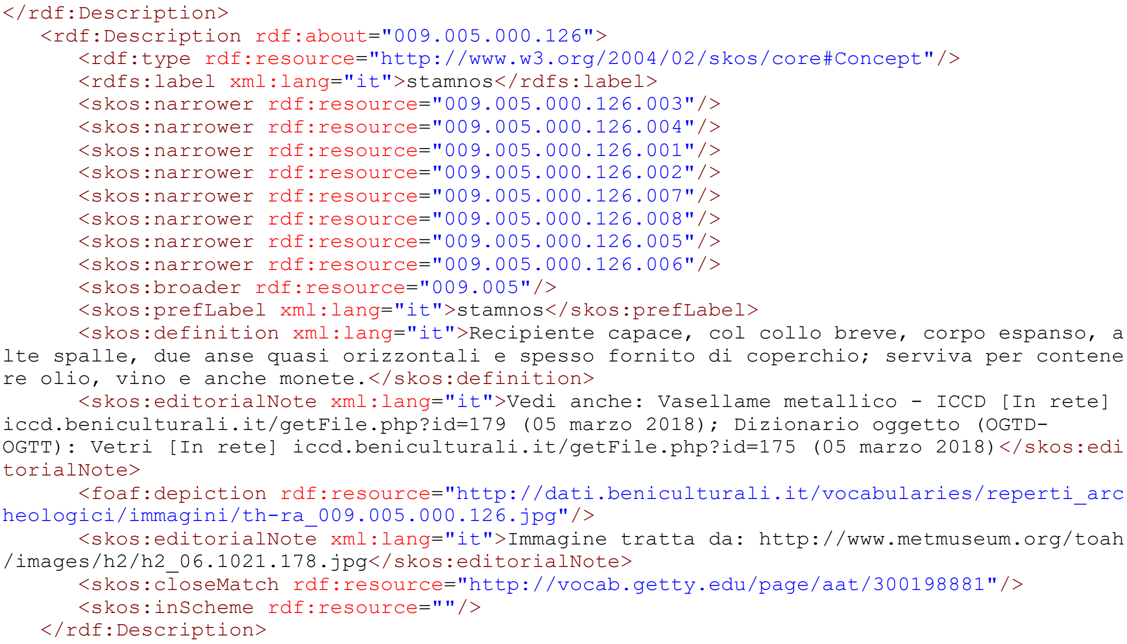
A manual evaluation revealed that one URI leads to a more generic term and the other one to a more specific term. For example the Italian term *letto* is linked both to the Getty AAT ‘Bed’ (generic) and to ‘Canopy Bed’ (specific). In these cases, instead of following the URI pointing to the specific reference, we choose to follow the most generic one, in accordance with the Italian term meaning. We opt for a manual evaluation due to the low presence of this phenomenon, but, alternatively, it could have been performed automatically making use of an external resource such as a dedicated dictionary.

Furthermore, the evaluation phase revealed a difference in the granularity of terms between the Italian ICCD Thesaurus and the other languages coming from the Getty AAT. Indeed, while the Italian terms result to be highly specific and fine-grained, many equivalents in the other languages are more in a relation of hyperonymity/hyponymity. For example, in the Italian Thesaurus there are several semantically and linguistically different types of relieves: their meanings change according to the following adjectives (e.g., *Rilievo + storico, funerario, votivo*, could be in English historical, funerary, votive + Relief). Nonetheless, the retrieved equivalent in English extracted from the Getty AAT is always ‘Relief’, as well as in Spanish is always ‘Relieve’ and in Dutch is ‘Reliëf’.

Finally, some terms in the different languages, as well as some definitions, are missing and we plan to implement the missing fields in the future. Table 1 shows the total number of terms for each language in the terminological database. Missing fields are due to data sparsity, since for each Italian term there are not always equivalent terms in

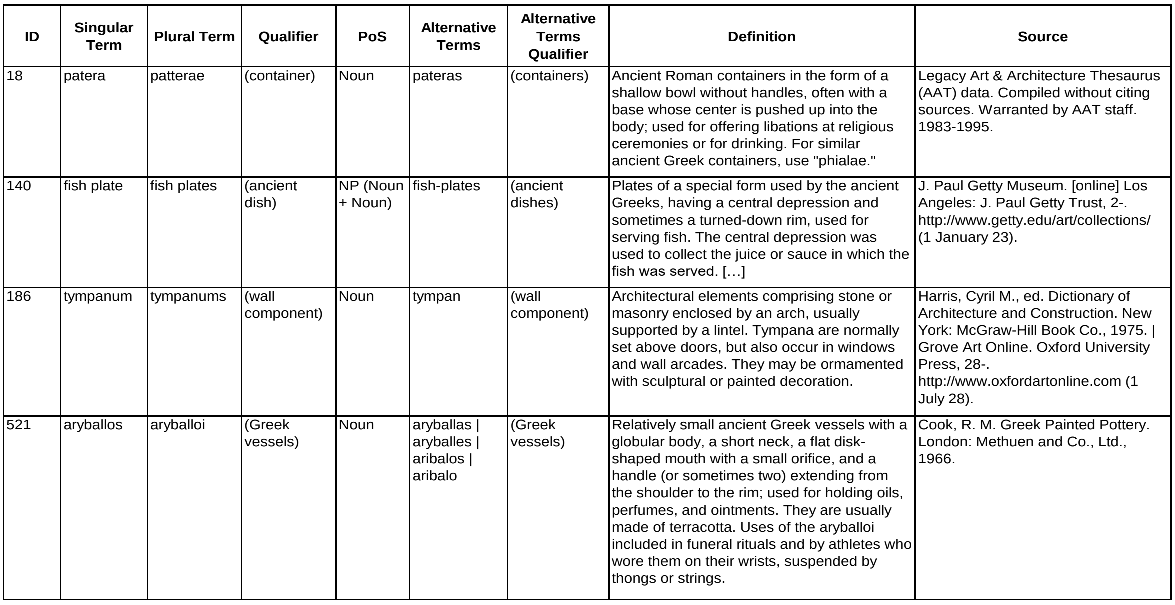
[18] https://drive.google.com/file/d/1cKvZPd6bdh7lrZ6plj1gGKatWvopqFo4/view The Glossary is released under Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)

[19] The ‘Qualifier’ field, enclosed between brackets, indicates the subfield the term belongs to, thus allowing the disambiguation in case of homographs (e.g. *Ax (weapon)* vs. *Ax (tool))*



```
</rdf:Description>
   <rdf:Description rdf:about="009.005.000.126">
      <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
      <rdfs:label xml:lang="it">stamnos</rdfs:label>
      <skos:narrower rdf:resource="009.005.000.126.003"/>
      <skos:narrower rdf:resource="009.005.000.126.004"/>
      <skos:narrower rdf:resource="009.005.000.126.001"/>
      <skos:narrower rdf:resource="009.005.000.126.002"/>
      <skos:narrower rdf:resource="009.005.000.126.007"/>
      <skos:narrower rdf:resource="009.005.000.126.008"/>
      <skos:narrower rdf:resource="009.005.000.126.005"/>
      <skos:narrower rdf:resource="009.005.000.126.006"/>
      <skos:broader rdf:resource="009.005"/>
      <skos:prefLabel xml:lang="it">stamnos</skos:prefLabel>
      <skos:definition xml:lang="it">Recipiente capace, col collo breve, corpo espanso, a
lte spalle, due anse quasi orizzontali e spesso fornito di coperchio; serviva per contene
re olio, vino e anche monete.</skos:definition>
      <skos:editorialNote xml:lang="it">Vedi anche: Vasellame metallico - ICCD [In rete]
iccd.beniculturali.it/getFile.php?id=179 (05 marzo 2018); Dizionario oggetto (OGTD-
OGTT): Vetri [In rete] iccd.beniculturali.it/getFile.php?id=175 (05 marzo 2018)</skos:edi
torialNote>
      <foaf:depiction rdf:resource="http://dati.beniculturali.it/vocabularies/reperti_arc
heologici/immagini/th-ra_009.005.000.126.jpg"/>
      <skos:editorialNote xml:lang="it">Immagine tratta da: http://www.metmuseum.org/toah
/images/h2/h2_06.1021.178.jpg</skos:editorialNote>
      <skos:closeMatch rdf:resource="http://vocab.getty.edu/page/aat/300198881"/>
      <skos:inScheme rdf:resource=""/>
   </rdf:Description>
```

Figure 1: Sample of the Italian term entry “stamnos” in the ICCD RDF/SKOS formalism.



ID	Singular Term	Plural Term	Qualifier	PoS	Alternative Terms	Alternative Terms Qualifier	Definition	Source
18	patera	patterae	(container)	Noun	pateras	(containers)	Ancient Roman containers in the form of a shallow bowl without handles, often with a base whose center is pushed up into the body; used for offering libations at religious ceremonies or for drinking. For similar ancient Greek containers, use "phialae."	Legacy Art & Architecture Thesaurus (AAT) data. Compiled without citing sources. Warranted by AAT staff. 1983-1995.
140	fish plate	fish plates	(ancient dish)	NP (Noun + Noun)	fish-plates	(ancient dishes)	Plates of a special form used by the ancient Greeks, having a central depression and sometimes a turned-down rim, used for serving fish. The central depression was used to collect the juice or sauce in which the fish was served. […]	J. Paul Getty Museum. [online] Los Angeles: J. Paul Getty Trust, 2-. http://www.getty.edu/art/collections/ (1 January 23).
186	tympanum	tympanums	(wall component)	Noun	tympan	(wall component)	Architectural elements comprising stone or masonry enclosed by an arch, usually supported by a lintel. Tympana are normally set above doors, but also occur in windows and wall arcades. They may be ornamented with sculptural or painted decoration.	Harris, Cyril M., ed. Dictionary of Architecture and Construction. New York: McGraw-Hill Book Co., 1975. | Grove Art Online. Oxford University Press, 28-. http://www.oxfordartonline.com (1 July 28).
521	aryballos	aryballoi	(Greek vessels)	Noun	aryballas | aryballes | aribalos | aribalo	(Greek vessels)	Relatively small ancient Greek vessels with a globular body, a short neck, a flat disk-shaped mouth with a small orifice, and a handle (or sometimes two) extending from the shoulder to the rim; used for holding oils, perfumes, and ointments. They are usually made of terracotta. Uses of the aryballoi included in funeral rituals and by athletes who wore them on their wrists, suspended by thongs or strings.	Cook, R. M. Greek Painted Pottery. London: Methuen and Co., Ltd., 1966.

Figure 2: Example of the English monolingual table.

all the other languages.

Language	Terms
Italian (IT)	1059
English (EN)	1026
Dutch (NL)	900
Spanish (ES)	593
German (DE)	376

Table 1: Number of terms for each language in the termbase.

4 Conclusions and future works

In this paper we present our Archaeo-Term Project aimed at the creation of a multilingual glossary on archaeology. The Glossary is the result of an extraction and merging process from two already available resources released according to the RDF Data Model, namely the RDF/SKOS version of the Italian ICCD Thesaurus and the LOD version of the multilingual Getty AAT.

The Archaeo-Term glossary is an ongoing project which will address, as future steps, the completion of missing data (terms, definitions, correspondences, examples, etc.) for English, Dutch, Spanish and German, as well as the enlargement of the glossary on the basis of the semi-automatic extraction of terminology from specialised corpora and other existing glossaries for the languages currently foreseen.

Furthermore, we also plan to implement the glossary with other languages such as French, Swedish, Chinese and Russian.

As future work we also plan to convert the result of Archaeo-Term project into more formalised formats, i.e., both TBX format (TermBase eXchange) to be used in connection with CAT-Tools and Ontolex-Lemon Model (McCrae et al., 2017), following the Linguistic Linked Open Data (LLOD) principles.

Finally, when we achieve a more complete version of the glossary we plan to publish it also on a Research Infrastructure Repository such as CLARIN.

Acknowledgments

This work has been partially supported by Programma Operativo Nazionale Ricerca e Innovazione 2014-2020 - Fondo Sociale Europeo, Azione I.2 “Attrazione e Mobilità Internazionale dei Ricercatori” Avviso D.D. n 407 del 27/02/2018 and by POR Campania FSE 2014-2020 “Dottorati di Ricerca a Caratterizzazione Industriale”.

We would like to thank Michele Stefanile for his support as expert in the domain of Archaeology. Authorship Attribution is as follows: Giulia Speranza is author of Section 2 and 3.2, Raffaele Manna is author of Section 3.1, Maria Pia di Buono is author of Section 1 and Johanna Monti is author of Section 3 and 4.

References

Maria Teresa Cabré. 1999. *Terminology: Theory, methods, and applications*, volume 1. John Benjamins Publishing.

Valentina Anita Carriero, Aldo Gangemi, Maria Letizia Mancinelli, Ludovica Marinucci, Andrea Giovanni Nuzzolese, Valentina Presutti, and Chiara Veninata. 2019a. Arco ontology network and lod on italian cultural heritage. In *ODOCH@ CAiSE*, pages 97–102.

Valentina Anita Carriero, Aldo Gangemi, Maria Letizia Mancinelli, Ludovica Marinucci, Andrea Giovanni Nuzzolese, Valentina Presutti, and Chiara Veninata. 2019b. Arco: The italian cultural heritage knowledge graph. In *International Semantic Web Conference*, pages 36–52. Springer.

Christian Chiarcos, John McCrae, Philipp Cimiano, and Christiane Fellbaum. 2013. Towards open data for linguistics: Linguistic linked data. In *New Trends of Research in Ontologies and Lexical Resources*, pages 7–25. Springer.

Michele Cortelazzo. 1994. Lingue speciali. *La dimensione verticale, Padova*.

David Crystal. 1997. The cambridge encyclopedia of language, wyd. 2. *New York*.

Pamela Faber and Clara Inés López Rodríguez. 2012. 2.1 terminology and specialized language. *A cognitive linguistics view of terminology and specialized language*, 20:9.

Achille Felicetti, Ilenia Galluccio, Cinzia Luddi, Maria Letizia Mancinelli, Tiziana Scarselli, and Antonio Davide Madonna. 2015. Integrating terminological tools and semantic archaeological information: the iccd ra schema and thesaurus. In *EMF-CRM@ TPDL*, pages 28–43.

Achille Felicetti, Daniel Williams, Ilenia Galluccio, Douglas Tudhope, and Franco Niccolucci. 2018. Nlp tools for knowledge extraction from italian archaeological free text. In *2018 3rd Digital Heritage International Congress (DigitalHERITAGE) held jointly with 2018 24th International Conference on Virtual Systems & Multimedia (VSMM 2018)*, pages 1–8. IEEE.

Maurizio Gotti. 2008. *Investigating specialized discourse*. Peter Lang.

John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The ontolex-lemon model: development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21.

Carlo Meghini, Roberto Scopigno, Julian Richards, Holly Wright, Guntram Geser, Sebastian Cuy, Johan Fihn, Bruno Fanini, Hella Hollander, Franco Niccolucci, et al. 2017. Ariadne: A research infrastructure for archaeology. *Journal on Computing and Cultural Heritage (JOCCH)*, 10(3):1–27.

Alan K Melby. 2012. Terminology in the age of multilingual corpora. *The Journal of Specialised Translation*, 18:7–29.

Alan Melby. 2015. Tbx: A terminology exchange format for the translation and localization industry. *201), Handbook of Terminology*, pages 393–424.

Sue Ellen Wright, Nathan Rasmussen, Alan K Melby, and L Warburton. 2010. Tbx glossary: a crosswalk between termbase and lexbase formats. In *Proceedings of developing, updating and coordinating technologies, dictionaries and lexicons for terminological consistency workshop*.