Revue de Traduction et Langues Volume21 Numéro1/2022
Journal of Translation Languages مجلة الترجمة واللغات
ISSN (Print) : 1112-3974 EISSN (Online) : 2600-6235

Le projet Archaeo-Term : premiers résultats

The Project Archaeo-Term: Initial Results

Johanna Monti
Université de Naples « L'Orientale » - Italie
jmonti@unior.it
UNIOR NLP Research Group
0000-0002-4563-5988
Maria Pia di Buono
Université de Naples « L'Orientale » - Italie
mpdibuono@unior.it
UNIOR NLP Research Group
0000-0001-7284-6556
Giulia Speranza
Université de Naples « L'Orientale » - Italie
gsperanza@unior.it
UNIOR NLP Research Group
0000-0002-9249-492X
Maria Centrella
Université de Naples « L'Orientale » - Italie
mcentrella@unior.it
0000-0002-4285-3963
Andrea De Carlo
Université de Naples « L'Orientale » - Italie
afdecarlo@unior.it
0000-0001-9116-8308

Comment citer cet article:
Monti, J., Di Buono, M. P., Speranza, G., Centrella, M., & De Carlo A. (2022). Le projet Archaeo-Term : premiers résultats. *Revue Traduction et Langues 21 (1)*, 121-136.

Reçu: 30/07/2022; **Accepté**: 21/08/2022, **Publié:** 31/08/2022

L'auteur correspondant : Monti Johanna

Keywords

Archaeology; Cultural Heritage; Multilingualism; Semantic Web; Terminological Resources; Thesaurus

Abstract

This article aims at describing the objectives, the theoretical and methodological background, the development, and the first results of the Archaeo-Term project of the University of Naples "L'Orientale", Department of Literary, Linguistic and Comparative Studies. The Archaeo-Term project has been developed within the YourTermCULT project promoted by the Terminology Without Borders Project of the Terminology Coordination Unit (TermCoord) of the European Parliament - Directorate-General for Translation (DGT) specifically for collecting terminology in different aspects related to culture. The aim of the Archaeo-Term project is to enhance the access to the archaeological data in several formats and languages. It represents a common effort to contribute to the creation of linguistic and terminological resources for the domain of Cultural Heritage (CH) and, in particular, for the sub-domain of archaeology, which is notably highly complex and fragmented. One of the first results of the Archaeo-Term project is the creation of a multilingual terminological resource for the domain of archaeology, which can be conveniently employed in different Natural Language Processing (NLP) tasks, including Machine Translation (MT). The first version of the Archaeo-Term multilingual terminological resource is available in 5 languages: Italian, English, Spanish, German, and Dutch and is publicly accessible online. With the objective of promoting a common and shared termbase across different languages, the Archaeo-Term terminological resource is addressed not only to a specialized audience such as experts in the field of archaeology but also as terminological support for translators and interpreters during their professional practice, as well as for a more general audience. The terminological resource is the result of an extraction and aggregation process carried out starting from two already existing thesauri: the Italian "Thesaurus per la definizione dei reperti archeologici" developed by the Italian Central Institute for Catalogue and Documentation (Istituto Centrale per il Catalogo e la Documentazione - ICCD) and the multilingual Art and Architecture Thesaurus (AAT) developed by the Getty Research Institute, which is among the most trustworthy and accurate resources in the domain of Cultural Heritage. Taking advantage of the Semantic Web formalisms applied to these terminological resources, we are able to extract and merge information from the aforementioned thesauri using SPARQL queries. Indeed, we run different queries against the SPARQL endpoint to enrich our multilingual terminological resource by extracting useful information about the different terminological entries. The information extracted and merged from these thesauri by means of several consecutive queries is as follows: the equivalent terms in the foreseen languages, the alternative terms and the plural forms, the domains and sub-domains, the definitions of the terms, and their sources. Furthermore, the extraction phase has been followed by an evaluation step aimed at checking missing information, verifying and adjusting possible misalignments among entries, and setting potential future implementations. As an ongoing project, we are also planning to enlarge the terminological resource with equivalent terms in other languages such as French, Swedish, Polish, Russian, and Chinese, with the aim of extending the language coverage also to non-European languages which are usually under-represented and low-resourced. As a first implementation with regards to the first version of the terminological

resource, we have currently collected: 1.059 entries in Italian, 1.055 in Spanish, 1.053 in English, 843 in Russian, 600 in Polish, 460 in German, 193 in French, and 82 in Chinese. To conclude, the Archaeo-Term project aims at promoting the creation of high-quality and trustworthy multilingual terminological resources for the domain of archaeology by also collaborating at the same time with institutions, experts in the field of terminology, linguistics, and cultural heritage.

Mots clés

Archéologie ; Patrimoine culturel ; Multilinguisme ; Web sémantique ; Ressources terminologiques ; Thesaurus

Résumé

Cet article vise à décrire les objectifs, les développements et les premiers résultats du projet Archaeo-Term, promu par l'Université de Naples "L'Orientale". Le projet a été développé dans le cadre du projet YourTermCULT, en collaboration avec le programme Terminologie sans frontières de l'Unité de coordination terminologique (TermCoord) du Parlement européen - Direction générale de la traduction (DG TRAD). L'objectif du projet Archaeo-Term est la création d'une ressource terminologique multilingue pour le traitement automatique des langues (TAL) dans le domaine de l'archéologie afin d'améliorer l'accès aux données archéologiques dans divers formats et langues. La première version de la ressource multilingue "Archaeo-Term Multilingual Glossary v1.0" (Speranza et al., 2020) contient des termes en 5 langues : italien, anglais, espagnol, allemand et néerlandais ; la deuxième version intègre le français, le suédois, le polonais, le russe et le chinois. Afin d'améliorer et de promouvoir l'utilisation d'une base terminologique commune entre différentes langues, le Glossaire multilingue de l'Archaeo-Term est conçu à la fois pour les utilisateurs spécialisés, tels que les chercheurs et les experts du domaine, et pour soutenir le travail des traducteurs et des interprètes, ainsi que pour un public général de non-spécialistes. La contribution présente le contexte théorique du projet, les données et la méthodologie utilisées pour développer Archaeo-Term, les résultats obtenus et leur évaluation et, enfin, les conclusions et les développements futurs du projet.

1. Introduction

Cet article vise à décrire les objectifs, les développements et les premiers résultats du projet Archaeo-Term, promu par l'Université de Naples « L'Orientale », Département d'études littéraires, linguistiques et comparées.

Le projet a été développé dans le cadre du projet YourTermCULT, en collaboration avec le programme Terminology Without Borders[1] de l'Unité de coordination terminologique (TermCoord)[2] du Parlement européen – Direction générale de la traduction (DG TRAD).

Parmi les différents projets promus par Terminology Without Borders (YourTermMED, YourTermENVI, YourTermTECH, YourTermFOOD, etc.),

[1] https://yourterm.eu/
[2] https://termcoord.eu/

YourTermCULT repose sur l'hypothèse que la culture est un domaine dans lequel la communication multilingue a besoin de messages clairs, adressés à des publics différenciés ; il a donc été conçu pour fonctionner dans certains des principaux aspects et secteurs de la culture (musées et institutions culturelles, archéologie, culture immatérielle, sport, musique).

L'objectif du projet Archaeo-Term est la création d'une ressource terminologique multilingue pour le traitement automatique des langues (TAL) dans le domaine de l'archéologie. Plus précisément, le projet Archaeo-Term vise à contribuer à l'amélioration de l'accès aux données archéologiques dans divers formats et langues.

La première version de la ressource multilingue « Archaeo-Term Multilingual Glossary v1.0 » (Speranza et al., 2020) renferme des termes en cinq langues : italien, anglais, espagnol, allemand et néerlandais, et en tant que recherche d'intérêt stratégique du Département d'études littéraires, linguistiques et comparées de l'Université de Naples « L'Orientale », prévoit de développer des ressources pour d'autres langues et en particulier pour le français, le suédois, le polonais, le russe et le chinois.

Afin d'améliorer et de promouvoir l'utilisation d'une base terminologique commune entre différentes langues, le Glossaire Multilingue d'Archaeo-Term est conçu tant pour les utilisateurs spécialisés, tels que les chercheurs et les experts dans le domaine, que pour soutenir le travail des traducteurs et des interprètes, ainsi que pour le grand public non spécialisé.

Une première version du Glossaire Multilingue est le résultat d'un processus d'extraction et de fusion de deux thésaurus (le Thesaurus pour la définition des découvertes archéologiques de l'*Istituto Centrale per il Catalogo e la Documentazione* – ICCD[3] et le *Thesaurus multilingue per l'Arte e l'Architettura* - ATT – du Getty Research Institute[4]), abondant en informations linguistiques, définitions, synonymes et sous-domaines.

Cette contribution présente le cadre théorique du projet, les données et la méthodologie utilisées pour développer Archaeo-Term, les résultats obtenus et leur évaluation, et enfin les conclusions et les développements à venir du projet.

2. Cadre théorique

La terminologie représente un domaine d'étude très complexe qui englobe différentes approches, principalement dans une perspective descriptive des données terminologiques. Comme l'ont souligné plusieurs chercheurs (Wright et al., 2010 ; Melby, 2012), la recherche en terminologie implique souvent l'utilisation de modèles et de formats de données très hétérogènes, ce qui implique que les ressources développées ne sont pas toujours faciles à réutiliser dans le cadre des technologies de traduction, telles

[3] http://www.iccd.beniculturali.it/
[4] https://www.getty.edu/research/tools/vocabularies/aat/

que les bases de données terminologiques et les mémoires de traduction, largement utilisées par les traducteurs professionnels.

Pour résoudre ce problème, un format standard a été développé, à savoir le TermBase eXchange (TBX) (Melby, 2015). Plus récemment, suite à la diffusion des technologies du Web sémantique, l'intérêt pour les ressources linguistiques informatiques et interopérables a augmenté, car elles constituent un élément essentiel du traitement des aspects sémantiques du langage naturel. En effet, les développeurs ont besoin de grandes collections de données linguistiques qui puissent être traitées par des algorithmes d'intelligence artificielle. Pour ce faire, plusieurs ressources alignées sur les principes de Linked Open Data (LOD) ont été publiées, en utilisant des formalismes tels que SKOS et Ontolex-Lemon, basés sur Resource Description Framework (RDF), pour représenter des glossaires, des vocabulaires et des taxonomies (Chiarcos et al., 2013).

Dans le domaine du patrimoine culturel, un certain nombre de ressources ont été développées ces dernières années. En particulier pour la langue italienne, l'ICCD a lancé en 2017 le projet ArCo[5] en collaboration avec l'Istituto di Scienze e Tecnologie della Cognizione (ISTC) du CNR pour que les données du Catalogue général du patrimoine culturel soient disponibles au format LOD (Carriero et al., 2019a ; Carriero et al., 2019b) et qu'elles soient accessibles, repérables et réutilisables aussi bien par les usagers que par des applications, et à même d'être connectées à d'autres ensembles de données disponibles au format LOD.

Une autre ressource, cette fois-ci multilingue, est l'Art & Architecture Thesaurus (AAT)[6], développé par le Getty Research Institute. Il s'agit d'un thésaurus multilingue utilisé pour la description des concepts liés à l'art, à l'architecture, aux arts décoratifs, à la culture matérielle et aux matériaux d'archives, accessible via un portail web ou via la version LOD (JSON, RDF, N3/Turtle, N-Triples), ainsi qu'au format XML et en tables relationnelles.

Un autre projet multilingue est l'iDAI.vocab[7], sous forme de vocabulaire contrôlé dans le domaine de l'archéologie développé pour plusieurs langues par le Deutschen Archäologischen Instituts (DAI). De plus, des ressources précieuses dans le domaine de l'archéologie ont récemment été développées dans le cadre de divers projets. Le projet ARIADNE (Meghini et al., 2017), notamment, propose un portail de données et de ressources pour pallier la fragmentation des différentes collections de données archéologiques. Plusieurs autres glossaires et thésaurus monolingues élaborés à des fins de catalogage sont également disponibles, tels que le FISH (Forum on Information Standards in Heritage)[8] au format LOD développé par Heritage Data[9] pour la langue anglaise ou les vocabulaires contrôlés déjà évoqués de l'*Istituto del Catalogo e La*

[5] http://wit.istc.cnr.it/arco/index.php?lang=it
[6] https://www.getty.edu/research/tools/%20vocabularies/aat/about.html
[7] https://idai.world/how/thesauri-and-controlled-vocabularies
[8] http://www.heritage-standards.org.uk/%20terminology/
[9] https://www.heritagedata.org/blog/%20vocabularies-provided/

Documentazione – ICCD. Il convient également de mentionner que certains glossaires sont publiés par des musées et des instituts culturels, comme le Thesaurus des noms d'objets du British Museum[10].

3. Données et méthodologie

Le point de départ de la création du Glossaire Multilingue d'Archaeo-Term est le *Thesaurus per la Definizione del bene reperti Archeologici* géré par l'*Istituto del Catalogo e La Documentazione* (ICCD). Ce thésaurus, initialement publié en version .pdf, a récemment été converti au format RDF/SKOS selon les formalismes du web sémantique et constitue l'une des bonnes pratiques adoptées par le Ministère de la culture (MiC) pour publier des informations institutionnelles sur le portail correspondant[11] conformément aux principes du Linked Open Data (LOD), dans le but de rendre les données et les informations produites par les institutions facilement repérables, réutilisables et librement accessibles.

Plus en détail, le format RDF (Resource Description Framework)[12] (Decker et al., 2000) est le modèle standard pour l'échange de données sur le web et permet d'exprimer des informations sur différents types de ressources, comprises comme des données, telles que des documents, des personnes, des objets physiques, des concepts. Le format RDF permet également de formaliser des informations dans un format *machine readable*, c'est-à-dire destiné à être traité par une machine ou une application, plutôt que par des utilisateurs humains. Le format RDF est également une norme qui sous-tend le web sémantique (Berners-Lee et al., 2001).

Le format SKOS (Simple Knowledge Organisation System)[13] (Miles et Bechhofer, 2009), quant à lui, permet de formaliser les ressources terminologiques, notamment les thésaurus et les taxonomies, ainsi que de partager et de relier les systèmes d'organisation des savoirs via le Web, ce qui permet le partage des données et des technologies entre différentes applications.

Le thésaurus ICCD dans sa version RDF/SKOS contient plus de 4 000 termes relatifs au domaine des découvertes archéologiques en italien, dont 1 059 termes ont un lien direct avec les termes correspondants dans d'autres langues présents dans la version LOD de l'Art and Architecture Thesaurus (AAT) du Getty Research Institute, qui représente – pour notre projet – la deuxième ressource où puiser principalement des données multilingues en rapport avec l'italien. L'AAT du Getty Research Institute, en effet, fait autorité dans le domaine du patrimoine culturel et représente l'un des thésaurus les plus fiables, offrant des termes dans de nombreuses langues.

Pour désigner une équivalence parfaite, SKOS prévoit l'utilisation de la propriété `skos:exactMatch` (correspondance exacte). La propriété `skos:closeMatch`,

[10] https://collectionstrust.org.uk/resource/british-museum-object-names-thesaurus/
[11] https://dati.cultura.gov.it/
[12] https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/#section-use-cases
[13] https://www.w3.org/2009/08/skos-reference/skos.html

quant à elle, désigne une correspondance entre des concepts qui ne sont pas parfaitement équivalents, mais plutôt une correspondance « proche ». Plus particulièrement, la propriété `skos:closeMatch` est une propriété que SKOS met à disposition pour effectuer des opérations de mappage et d'alignement. En effet, elle sert à relier deux concepts similaires qui peuvent être utilisés de manière interchangeable mais qui appartiennent à deux ressources différentes.

Dans le thésaurus ICCD, cette propriété a été utilisée, par exemple, pour relier l'entrée italienne « acroterio a disco » à l'entrée multilingue « acroterio » dans l'AAT. Les deux entrées n'ont pas de correspondance parfaite, ni exacte, mais plutôt « proche » dans la mesure où l'entrée italienne est plus spécifique que l'entrée multilingue. Ainsi, via la propriété `skos:closeMatch`, chacun des 1 059 termes du thésaurus ICCD est lié à un « Universal Resource Identifier » (URI) qui identifie de manière unique les termes dans l'AAT du Getty. Il faut noter à cet égard que le travail de mise en correspondance entre les termes du thésaurus ICCD et les termes de l'AAT du Getty Research Institute a été réalisé dans le cadre du projet ARIADNE (Felicetti et al., 2015), en utilisant uniquement la propriété `skos:closeMatch`.

```
<rdf:Description rdf:about="003.002.001.001">
      <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
      <rdfs:label xml:lang="it">acroterio</rdfs:label>
      <skos:narrower rdf:resource="003.002.001.001.005"/>
      <skos:narrower rdf:resource="003.002.001.001.006"/>
      <skos:narrower rdf:resource="003.002.001.001.003"/>
      <skos:narrower rdf:resource="003.002.001.001.004"/>
      <skos:narrower rdf:resource="003.002.001.001.007"/>
      <skos:narrower rdf:resource="003.002.001.001.008"/>
      <skos:broader rdf:resource="003.002.001"/>
      <skos:narrower rdf:resource="003.002.001.001.001"/>
      <skos:narrower rdf:resource="003.002.001.001.002"/>
      <skos:prefLabel xml:lang="it">acroterio</skos:prefLabel>
      <skos:editorialNote xml:lang="it">Immagine tratta da: Adam, Jean-Pierre,
L'arte di costruire presso i romani: materiali e tecniche, Milano: Longanaesi,
1984, p. 359</skos:editorialNote>
      <skos:definition xml:lang="it">Elemento ornamentale posto al vertice ed alle
estremità del frontone del tempio. Gli acroteri possono essere costituiti da
elementi decorativi come la voluta e da figure a tutto tondo come la
sfinge.</skos:definition>
      <skos:editorialNote xml:lang="it">Fonte: [EAA, vol. I,
p.15</skos:editorialNote>
      <foaf:depiction
rdf:resource="http://dati.beniculturali.it/vocabularies/reperti_archeologici/immagi
ni/th-ra_003.002.001.001.jpg"/>
      <skos:closeMatch rdf:resource="http://vocab.getty.edu/page/aat/300002214"/>
      <skos:inScheme rdf:resource=""/>
```

Figure 1. Formalisation RDF/SKOS de l'entrée « acroterio » dans le thésaurus ICCD

Available online online at https://www.asjp.cerist.dz/en/Articles/155

La figure 1 montre l'entrée formalisée RDF/SKOS « acroterio » dans le thésaurus ICCD pour la définition « Reperti Archeologici ». Comme on peut le voir, le terme est représenté avec la propriété `skos:prefLabel` et l'indication de la langue (`xml:lang= "it"`). La définition du terme est exprimée via la propriété `skos:definition` et la source via la propriété `skos:editorialNote`. La propriété `skos:closeMatch` permet enfin d'indiquer l'URI relatif au terme correspondant dans l'AAT du Getty Research Institute. La propriété `skos:broader`, quant à elle, désigne la macro-catégorie et les micro-catégories au sein de la taxonomie conceptuelle hiérarchique à laquelle le terme appartient. Cette hiérarchie est exprimée par des codes numériques à trois chiffres. Par exemple, le terme « acroterio » appartient à la macro-catégorie « EDILIZIA » identifiée par le code `003`, et en particulier aux micro-catégories « ELEMENTI ARCHITETTONICI » par le code `002`, et « ELEMENTI DECORATIVI E DI ARREDO » identifiée par le code `001` (`<skos:borader rdf:resource= "003.002.001"/>`).

Les URI du thésaurus ICCD renvoyant à l'AAT nous permettent de récupérer les informations dont nous avons besoin pour structurer notre ressource terminologique dans une perspective multilingue. Notre approche méthodologique est basée sur l'interrogation du SPARQL Endpoint[14] du Getty Research Institute pour accéder aux informations sur les termes en formulant diverses *queries* (requêtes).

Les informations à récupérer et à associer aux entrées de notre ressource sont : les termes équivalents dans les différentes langues, les termes alternatifs et les formes plurielles possibles, les définitions des termes et les sources. Plusieurs requêtes consécutives avec des objectifs différents ont été nécessaires pour cette opération afin d'extraire le plus d'informations possible, utiles pour l'enrichissement de notre ressource multilingue.

Pour commencer, une requête a été mise en place pour analyser le thésaurus ICCD et lire chaque URI faisant référence au terme correspondant dans l'AAT. Les termes du TCA sont formalisés et représentés conformément à la propriété `skos:prefLabel`, les termes alternatifs et les formes au singulier, quant à eux, sont représentés avec la propriété `skos:altLabel`. Ces deux propriétés consistent en une valeur lexicale et en une étiquette linguistique, associée à la valeur lexicale, pour chaque URI. Il convient de souligner que les termes dits préférentiels (`prefLabel`) dans le TCA sont tous au pluriel, tandis que la section sur les termes alternatifs (`altLabel`) énumère les termes au singulier. Cela semble être une particularité d'organisation et de gestion des thésaurus du Getty Research Institute.

Par ailleurs, l'objectif étant d'extraire les termes correspondants dans différentes langues, une requête supplémentaire a été mise en place pour extraire les termes équivalents dans les langues respectives avec les étiquettes linguistiques correspondantes.

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

La même opération a été également répétée pour extraire les termes alternatifs dans les différentes langues avec les étiquettes linguistiques correspondantes.

Une autre requête a ensuite été formulée afin de lire les URI et de recueillir les définitions de chaque terme et les sources d'où proviennent ces définitions, ainsi que leurs étiquettes linguistiques, toutes deux définies via la propriété `skos:scopeNote`.

Enfin, les sous-domaines auxquels appartiennent les termes ont également été extraits. Pour les termes italiens, les sous-domaines ou catégories sont extraits du thésaurus ICCD, tandis que pour toutes les autres langues, les sous-domaines proviennent de l'AAT. Nous considèrons que l'explication des sous-domaines est essentielle pour désambiguïser les cas éventuels d'homographie comme, par exemple, dans le cas du terme anglais *ax* (hache) qui peut appartenir à la fois aux sous-domaines des armes (*weapon*) et des outils (*tool*). L'indication du sous-domaine est également particulièrement utile dans les cas où des langues différentes utilisent des représentations linguistiques distinctes pour deux termes appartenant à des sous-domaines différents.

La figure 2 montre la version RDF du TAA en se concentrant sur l'entrée « acroteria », et les différents équivalents multilingues, y compris les formes alternatives, correspondant à l'italien « acroterio » auquel on accède via l'URI exprimé par la propriété `skos:closeMatch` dans le thésaurus ICCD. Comme on peut le constater, le terme privilégié dans le TCA pour chacune des langues est au pluriel : *acroteria* pour l'anglais (en) ; tandis que la forme au singulier est présente comme terme alternatif : *acroterion* pour l'anglais (en). De plus, il existe d'autres termes alternatifs pour l'anglais comme *acroterium* et *acroters.*

Ainsi, les informations sur les termes dans les différentes langues ne sont pas homogènes. Elles peuvent varier et, dans certains cas, peuvent même être absentes. Par exemple, l'équivalent français du terme italien « acroterio » ou le `skos:altlabel` pour le néerlandais est absent dans l'AAT.

Available online online at https://www.asjp.cerist.dz/en/Articles/155

```
<skos:prefLabel xml:lang="en">acroteria</skos:prefLabel>
<skos:prefLabel xml:lang="zh-hant">平底三角牆頂飾底</skos:prefLabel>
<skos:prefLabel xml:lang="nl">acroteriën</skos:prefLabel>
<skos:prefLabel xml:lang="de">Akroterion</skos:prefLabel>
<skos:prefLabel xml:lang="pt">acrotério</skos:prefLabel>
<skos:prefLabel xml:lang="es">acróteras</skos:prefLabel>
<skos:altLabel xml:lang="en">acroterion</skos:altLabel>
<skos:altLabel xml:lang="en">acroterium</skos:altLabel>
<skos:altLabel xml:lang="en">acroters</skos:altLabel>
<skos:altLabel xml:lang="zh-hant">三角楣頂屋頂雕像</skos:altLabel>
<skos:altLabel xml:lang="zh-hant">三角楣頂雕像底座</skos:altLabel>
<skos:altLabel xml:lang="zh-hant">雕刻飾物</skos:altLabel>
<skos:altLabel xml:lang="zh-hant">脊頭</skos:altLabel>
<skos:altLabel xml:lang="zh-hant">山牆飾物底座</skos:altLabel>
<skos:altLabel xml:lang="de">Akroterien</skos:altLabel>
<skos:altLabel xml:lang="de">Akroter</skos:altLabel>
<skos:altLabel xml:lang="de">Akrotere</skos:altLabel>
<skos:altLabel xml:lang="de">Akroteren</skos:altLabel>
<skos:altLabel xml:lang="de">Akroteria</skos:altLabel>
<skos:altLabel xml:lang="es">acrótera</skos:altLabel>
<skos:altLabel xml:lang="es">acroterio</skos:altLabel>
<skos:altLabel xml:lang="es">acrotera</skos:altLabel>
```

Figure 2. Formalisation par le RDF de l'entrée « acroteria » dans l'AAT

Une fois que toutes les informations recherchées et disponibles sur les termes ont été extraites, les URI ont été remplacés par des ID numériques uniques afin de fournir un code d'identification pour chaque entrée de notre ressource.

Enfin, la ressource a été organisée dans un fichier tabulaire au format .xlsx composé de tableaux (ou fiches) monolingues individuels pour chaque langue fournie (Figure 3) et d'un tableau synoptique de synthèse multilingue contenant les termes au singulier dans toutes les langues (Figure 4).

ID	SINGULAR TERM	PLURAL TERM	QUALIFIER	PoS	ALTERNATIVE TERM	ALT. TERM QUALIFIER	DEFINITION	SOURCE
795	bed	beds	(furniture)	N	bedstocks		Generally, the sleeping places	Legacy Art & Arc
796	candlestick	socket candleholder		N	socket candlesticks		Candleholders with a single ca	Legacy Art & Arc
797	dipper	dippers	(serving utensil)	N			Utensils consisting of concave	Legacy Art & Arc
798	casket	caskets	(personal gear)	N			Small chests or boxes for holdi	Legacy Art & Arc
799	sail	sails	(equipment)	N			Sheets or panels of material, d	Legacy Art & Arc
800	loom	looms	(textile tool)	N			Frames, devices, or machines	Becker, John. Pa
801	joint	joints	(connection)	N	connections		The apparatus used or the ma	Legacy Art & Arc
802	atlas	atlases	(supporting element)	N	telamon | telamones | telamons		Male figures, usually nude or c	
803	capital	capitals	(column component)	N	chapiters		The uppermost members of co	Dizionario encicl
804	harpoon	harpoons		N			Hunting weapons consisting of	Legacy Art & Arc
805	mirror	mirrors		N			Objects with a highly polished	Legacy Art & Arc
806	razor	razors		N			Sharp-edged instruments, mar	Legacy Art & Arc
807	fountain	fountains		N			Structures with apertures desig	Legacy Art & Arc
808	lantern	lanterns	(lighting device)	N	lanthoms | lanthorns		Lighting devices, fixed or porta	Hough, Walter. C
809	pin	pins	(jewelry)	N			Ornaments consisting essentia	Random House
810	chain	chains	(object genre)	N			Series of objects connected or	Random House
811	cowry shell	cowrie shells		NP (N+N)	courie shell | cowrie shell		Shell of any of numerous marir	Webster's Third
812	mascaron	mascarons		N	grotesque masks | mascaroons	(motifs)	Motifs representing grotesque	Legacy Art & Arc
813	cement clinker	cement clinkers		NP (N+N)	clinker		Partially fused product from a k	
814	intarsio	intarsios	(process)	N	tarsia | incrustatio | loricatio | in	(process)	Decorative wood process in wh	Grove Art Online
815	sphere-shaped	spheres-shaped		NP (N+V_PP)	spherical		Having the shape or form of a	Oxford English D

Figure 3. Tableau monolingue pour la langue anglaise

ID	IT	EN	ES	NL	DE
1	corona funeraria	funeral ornament			
2	balsamario	unguentarium		unguentarium	
3	lacrimatoio	unguentarium		unguentarium	
4	unguentario	unguentarium		unguentarium	
5	cintura per la sospensione delle armi	cartridge belt	canana	patroongordel	
6	trozzella	nestoride			
7	collana	necklace	collar	halsketting	Halskette
8	parazonio	left-hand dagger	daga de zurdo	pareerdolk	
9	colonna sepolcrale	grave pillar		grafzuil	
10	portabrace	coal hod	caja de carbón	kolenkit	

Figure 4. Tableau synoptique multilingue pour l'italien, l'anglais, l'espagnol, le néerlandais et l'allemand

Pour ce qui est des tableaux monolingues, toutes les données récupérées ont été automatiquement classées dans des tableaux distincts en fonction du code de la langue ou du positionnement pour chaque entrée terminologique (p. ex. `it` pour l'italien, `en` pour l'anglais, `fr` pour le français, etc.). Il est à noter que les informations (terme au singulier, définition, source de la définition, *qualifier* – c'est-à-dire domaine) relatives à la langue italienne proviennent du Thésaurus monolingue ICCD, tandis que les mêmes informations relatives aux termes des autres langues proviennent de l'AAT.

En ce qui concerne le tableau synoptique multilingue, les termes dans les différentes langues ont été alignés en fonction de l'identifiant commun attribué précédemment. Par exemple, l'identifiant numérique 431 désignera le terme « acroterio » dans toutes les langues de la ressource terminologique et, par conséquent, dans les tableaux monolingues individuels.

La ressource terminologique ainsi développée a en outre été mise en ligne[15] en format ouvert pour les langues italienne, anglaise, allemande et néerlandaise, en utilisant la plateforme code source ouvert Lexonomy (Měchura, 2017) pour la création de dictionnaires électroniques en ligne. La figure 5 montre l'entrée « acroterio » désignant la partie du discours (noun), soit le substantif, les sous-domaines auxquels le terme appartient, la définition du terme, la source de la définition et, enfin, les équivalents terminologiques récupérés dans l'AAT du Getty Research Institute dans les différentes langues fournies, lorsqu'ils sont disponibles.

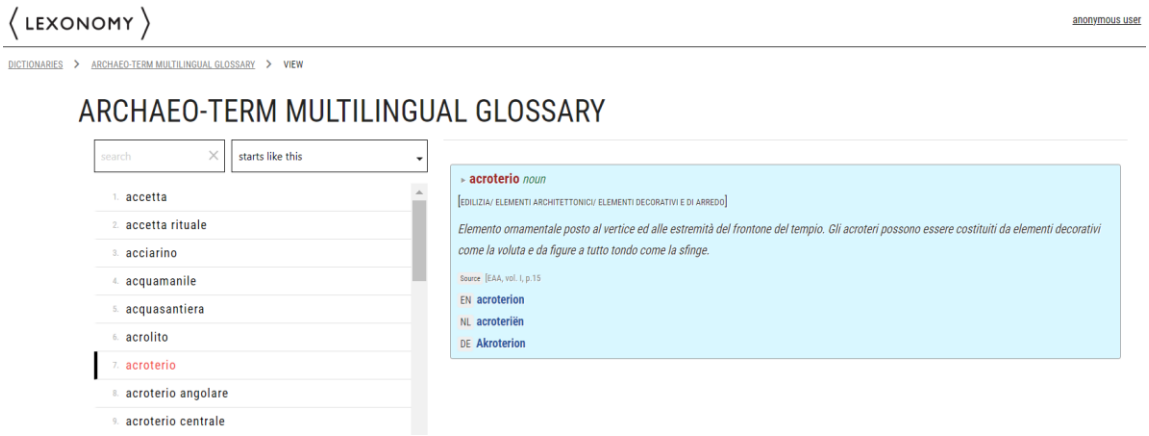

Figure 5. Exemple d'une entrée du Glossaire Multilingue Archaeo-Term dans Lexonomy

4. Résultats et évaluation

Le Glossaire Multilingue, dans sa première version réalisée à l'aide de techniques d'extraction automatique, se compose de :

1 059 entrées en italien,

[15] https://www.lexonomy.eu/#archaeo-term

1026 en anglais,
900 en néerlandais,
593 en espagnol
376 en allemand.

La phase d'extraction a été suivie d'une phase d'évaluation pour vérifier l'exhaustivité et la correspondance des informations assemblées et les éventuels problèmes à résoudre à l'avenir. Par exemple, dans le cas de 9 termes italiens, la présence de deux équivalents anglais a été notée, c'est-à-dire deux URI `skos:closeMatch` liés à l'AAT Getty, au lieu d'un seul. Un examen plus approfondi a révélé qu'un URI pointait vers un terme plus générique, l'autre URI, quant à lui, pointait vers un terme plus spécifique, comme dans le cas du terme italien « letto » (lit) qui était lié à l'AAT tant par le terme générique équivalent « bed » que par le terme « canopy bed » (lit à baldaquin), qui était plus spécifique. Dans ces cas, nous avons opté pour un nettoyage manuel, dicté par la faible fréquence du phénomène, en ne conservant que le lien vers le terme le plus générique, conformément à l'équivalent italien original. Cependant, la phase de vérification peut également être réalisée de manière automatique, en utilisant une ressource externe comme, par exemple, un glossaire ou un dictionnaire dédié.

En outre, la phase d'évaluation a également révélé un niveau différent de « granularité » et de spécificité des termes entre les deux thésaurus mappés. En effet, alors que le Thésaurus italien de l'ICCD présente des termes spécifiques et très détaillés, l'AAT du Getty Research Institute présente des termes plus génériques pour les langues sélectionnées. Par conséquent, de nombreuses relations entre les termes dans différentes langues s'avèrent être de type hyperonymiques/hyponymiques. Par exemple, le Thesaurus ICCD comprend différents types de « rilievo » (relief) : *rilievo storico*, *rilievo funerario*, *rilievo votivo* ; dans l'AAT, cependant, cette variété terminologique et sémantique n'est pas disponible, de sorte que l'équivalent anglais – et par conséquent dans les différentes langues du Thesaurus – pour les divers types de « rilievo » est toujours le terme *relieve*, qui s'avère être un hyperonyme générique par rapport à l'original italien. La recherche, également par le recours à des sources externes, d'une plus grande équivalence entre les termes est l'une des activités que nous prévoyons de poursuivre dans la mise en œuvre de la ressource terminologique.

Enfin, comme mentionné plus haut, il n'a pas toujours été possible d'extraire toutes les informations dans les différentes langues en raison des inhomogénéités et des lacunes de l'AAT. En effet, certains termes italiens n'ont pas d'équivalents dans certaines langues, ou dans certaines définitions, c'est pourquoi un enrichissement à partir d'autres ressources externes est prévu à l'avenir.

5. Conclusion

Le projet Archaeo-Term représente une première tentative et un effort conjoint pour contribuer à la création de ressources terminologiques et linguistiques pour un domaine fragmenté et complexe tel que le patrimoine culturel, en particulier pour le domaine de l'archéologie.

Plus précisément, la création du Glossaire Multilingue vise à inclure non seulement les langues européennes, mais aussi à étendre sa portée linguistique aux langues non européennes et aux langues peu représentées, pour lesquelles il est bien connu que les ressources spécialisées sont rares.

Actuellement, le projet est encore en phase de croissance et d'expansion, et nous travaillons à la mise en œuvre manuelle des termes manquants et à l'inclusion de termes équivalents dans d'autres langues, dans l'intention de fournir également des équivalents plus précis entre les différentes langues dans les cas de « granularité » différente, en s'appuyant également sur des sources supplémentaires comme, par exemple, des corpus spécialisés pour l'extraction de la terminologie du domaine. En tant que premier chiffre provisoire, encore en cours d'enrichissement, par rapport à la première version du glossaire, sont en effet rédigées :

1 059 entrées en italien,
1 055 en espagnol,
1 053 en anglais,
843 en russe,
600 en polonais,
460 en allemand,
193 en français,
82 en chinois.

Le projet Archaeo-Term se veut donc un promoteur de ressources linguistiques multilingues fiables et de qualité pour le domaine de l'archéologie grâce à une coopération étroite entre les institutions, les experts en terminologie et en linguistique et les experts du domaine. La ressource terminologique créée est indispensable pour les linguistes, les traducteurs et les interprètes, pour les acteurs du domaine du patrimoine culturel ainsi qu'extrêmement utile dans diverses applications de traitement automatique des langues.

Available online online at https://www.asjp.cerist.dz/en/Articles/155

Références

[1] Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. Scientific American, 284(5), 34-43.

[2] Carriero, V. A., Gangemi, A., Mancinelli, M. L., Marinucci, L., Nuzzolese, A. G., Presutti, V., & Veninata, C. (2019a, October). ArCo: The Italian cultural heritage knowledge graph. In International Semantic Web Conference, 36-52. Springer, Cham.

[3] Carriero, V. A., Gangemi, A., Mancinelli, M. L., Marinucci, L., Nuzzolese, A. G., Presutti, V., & Veninata, C. (2019b, June). ArCo ontology network and LOD on Italian Cultural Heritage. In ODOCH@ CAiSE, 97-102.

[4] Chiarcos, C., McCrae, J., Cimiano, P., and Fellbaum, C. (2013). Towards open data for linguistics: Linguistic linked data. In New Trends of Research in Ontologies and Lexical Resources, 7-25. Springer.

[5] Decker, S., Melnik, S., Van Harmelen, F., Fensel, D., Klein, M., Broekstra, J. and Horrocks, I. (2000). The semantic web : The roles of XML and RDF. IEEE Internet computing, 4(5), 63-73.

[6] Felicetti, A., Galluccio, I., Luddi, C., Mancinelli, M.L., Scarselli, T., and Madonna, A.D. (2015). Integrating terminological tools and semantic archaeological information : The ICCD RA schema and thesaurus. In EMFCRM@ TPDL, 28-43.

[7] Měchura, M. B. (2017, September). Introducing Lexonomy : an open-source dictionary writing and publishing system. In Electronic Lexicography in the 21st Century : Lexicography from Scratch. Proceedings of the eLex 2017 conference, 19-21.

[8] Meghini, C., Scopigno, R., Richards, J., Wright, H., Geser, G., Cuy, S., & Vlachidis, A. (2017). ARIADNE : A research infrastructure for archaeology. Journal on Computing and Cultural Heritage (JOCCH), 10(3), 1-27.

[9] Miles, A., & Bechhofer, S. (2009). SKOS simple knowledge organization system reference. W3C recommendation, 18, W3C.

[10] Speranza, G., Manna, R., di Buono, M. P., and Monti, J. (2020). The Archaeo-Term Project : Multilingual Terminology in Archaeology. In CLiC-it 2020 Italian Conference on Computational Linguistics 2020. CEUR.

[11] Melby, A. K. (2012). Terminology in the age of multilingual corpora. The Journal of Specialised Translation, 18 :7-29.

[12] Melby, A. K. (2015). TBX : A terminology exchange format for the translation and localization industry. 201), Handbook of Terminology, 393-424.

[13] Wright, S. E., Rasmussen, N., Melby, A. K., & Warburton, L. (2010, October). TBX glossary: a crosswalk between termbase and lexbase formats. In Proceedings of developing, updating and coordinating technologies, dictionaries and lexicons for terminological consistency workshop.

Remerciements
Nous tenons à remercier le réseau LTT pour l'organisation des 12e Journées du Réseau LTT en 2021. Nous tenons également à remercier l'équipe du CERIST qui héberge la revue TRANSLANG sur ASJP.

Notice bio-bibliographique des auteurs

Johanna Monti est Professeure associée en didactique des langues étrangères au département de littérature, linguistique et études comparées à l'université de Naples-L'Orientale, elle est coordinatrice de la formation de Master en traduction et linguistique computationnelle.

Maria Pia di Buono est doctorante et maitre assistante en linguistique générale au département des sciences sociales, politiques et communication à l'Université de Salerne, son domaine de recherche s'articule autour de la terminologie, la gestion des connaissances et à la sémantique ontologique

Giulia Speranza doctorante à l'université de Naples-L'Orientale au département de linguistique littéraire et études comparatives, son domaine de recherche s'articule autour de la translinguistique, données archéologiques et communication internationale, elle mène ses recherches sous la direction de Johanna Monti.

Centrella Maria, enseignante au département des études littéraires et linguistique comparative à l'université de Naples-L'Orientale.

Andrea De Carlo est enseignant de Philologie Polonaise à l'université de Naples-L'Orientale. De 2006 à 2011, il a été professeur contractuel de langue et de littérature polonaises à l'Université de Salento (Lecce, Italie), et pendant l'année universitaire 2020-2021 à l'Université de Bari Aldo Moro. En 2010, il a obtenu un doctorat de l'Université de Salento avec une thèse sur Dante dans la Pologne du XIXe siècle. Ses recherches portent sur la littérature polonaise, les relations culturelles entre l'Italie et la Pologne et la traduction poétique. Il travaille actuellement sur l'édition critique de La Comédie Divine traduite par J.I. Kraszewski. Il a reçu la qualification scientifique nationale de professeur associé en slavistique. Il est membre de l'AIS (Association italienne des slavistes) et membre du conseil d'administration de l'AIP (Association italienne des polonistes). Son domaine de recherche s'articule autour de la traduction poétique et des relations culturelles entre l'Italie et la Pologne, ses travaux sont régulièrement publiés dans des revues académiques italiennes et polonaises.