Tedi: a Platform for Ontologisation of Multilingual Terminologies for the Semantic Web

A Use Case from the Domain of Ancient Greek Cultural Heritage

Maria Papadopoulou\textsuperscript{1,2}, Christophe Roche\textsuperscript{1,2}

1) Equipe Condillac « Terminology & Ontology » - Listic University Savoie Mont-Blanc, France
2) Knowledge Engineering & Terminology Research Centre University of Liaocheng, China
E-mail: firstname.lastname@univ-savoie.fr

Abstract- The vision of the Semantic Web is machine understandability for all data currently stored in web-based resources. Terminological resources, which follow the ISO (International Organization for Standardization) standards on terminology in defining concepts as unique combinations of essential characteristics (ISO 1087-1), need to become computable and Semantic Web compliant. This paper, first, describes the theoretical approach and the tool-assisted method, which underlies the turning of these terminologies into Semantic web compliant ontologies. Next, this paper presents Tedi (ontoTerminology editor), the platform developed for building multilingual terminologies, which share the same formal domain ontology. Tedi allows to export these terminologies into OWL (Web Ontology Language), RDF (Resource Description Framework), JSON (JavaScript Object Notation), and in a number of other formats, including multilingual HTML (Hyper Text Markup Language) electronic dictionaries of terms. Tedi is based on a theory of concept dedicated to Terminology. Semantics is defined as the relation between terms (natural language units with meaning specialized to a domain of knowledge) and concepts (units of thought whose meaning is formally expressed as a set of essential characteristics), according to the discipline of Terminology. Tedi stores the linguistic and the conceptual dimensions in two related, yet distinct systems. This formal theory, which supplies the semantic onto-terminological layer needed for deeper data interpretability by machines, is less contrived and far more intuitive to use. It empowers domain experts to build their own semantic multilingual terminological dictionaries without having to be aware of logical formalisms like description logics. Semantic content management systems are direly needed in the domain of ancient Greek cultural heritage. The remainder of the paper will illustrate this particular point with a use case from the domain of ancient Greek dress terminology presented from the point of view of the user (domain expert).

Keywords- formal domain ontology; multilingual terminologies; ISO (and W3C (World Wide Web Consortium) standards; Tedi (ontoTerminology editor) software platform; ancient Greek cultural heritage.

I. INTRODUCTION

This paper proposes a tool-assisted method to design and create multilingual domain ontotermologies (i.e., terminologies whose conceptual system is a formal domain ontology) relying on a definition of concept as a set of essential characteristics. An essential characteristic is such that, if removed from the object, the object is no more what it is, e.g., mortal for ‘human being’). Such ontotermologies are both ontologies that represent and model the concepts of a domain of specialized knowledge and terminologies that capture the verbal expression of this knowledge in different natural languages. The approach is based on the assumption that the same conceptualization of a specific domain can be shared across different linguistic communities, albeit expressed differently due to the difference in the linguistic medium. It follows that a formal domain ontotermology can be built in order to capture a/ the conceptual layer of the domain of interest, and b/ the multilingual sets of terms denoting the concepts in the ontology. The concepts are defined in an artificial and formal language embedded in a user-friendly interface. The definitions of terms in natural language are built from the formal definitions of the concept each term denotes. This permits to guarantee some logical properties, such as coherence and completeness. What this achieves is a degree of standardization necessary for verbal communication among experts, inside and across communities of practice, based on a common understanding of their domain. This opens up new perspectives for the operationalization of terminologies for IT (Information Technology) applications. The approach is extremely useful for solving the problem of how to describe object-based knowledge of a part of the world in relation to the textual resources that refer to the same part, as is often the case in archaeology, classics, and cultural heritage studies.

The remainder of this paper is organized as follows: Section II describes the motivation that led to building the ontotermology editor Tedi [1]. Section III presents related work and briefly explains why it is not sufficient. Section IV addresses the theoretical underpinnings of the ontotermology approach. Section V describes the Tedi platform in terms of interfaces and details a use case from the domain of application. Our domain of choice was Greek dress, a domain which is deep-seated in modern perceptions of ancient Greek culture. The conclusions and future work section closes the article.

II. MOTIVATION

Ontologies and terminologies are at the core of the Semantic Web [2]. Ontologies, defined as “an explicit specification of a conceptualization” [3] mainly rely on description logics for their knowledge theory and on W3C
interchange formats for their formal representation [4]. The dominant formalism for representing ontology has been the T-Box (assertions on concepts) and A-Box (assertions on individuals) in Description Logics (DL) (alias terminological logics [5]). “Concepts represent sets of individuals, roles represent binary relations between the individuals, and individual names represent single individuals in the domain. Readers familiar with first-order logic will recognize these as unary predicates, binary predicates, and constants” [6]. Readers with no such background, however, will have difficulty grappling with the notion of DLs (Description Logics), better known for their decidability and the ability to infer additional knowledge, than for being intuitive [7]. The most popular free open-source editor for authoring ontologies based on these principles is Protégé [8] thanks to its powerful functionalities.

Not all terminologies rely on description logics for their conceptual system. Some terminologies follow the principles of the ISO standards for terminology work, which better match the way domain experts reason, because they are less contrived. There are numerous ISO standards for terminology work and no counterparts for dealing with ontology. The ISO 1087-1 and ISO 704, the standards on which all others should rely, were designed in times when the vision of the Semantic Web was not yet on the horizon (for a brief historical account see [9]). Their single goal was communication between humans, not IT applications [10], this is why they should be revised [11].

ISO 1087-1 [12] defines Terminology a/ as the “science studying the structure, formation, development, usage and management of terminologies in various subject fields”, and b/ as the result of the application of this science to a dedicated specialized domain, i.e., a “set of designations belonging to one special language”. ISO 1087-1 defines concept as a “unit of knowledge created by a unique combination of characteristics” and term as a “verbal designation of a general concept in a specific subject field”. Representing concepts as sets of essential characteristics, not as sets of individuals (which is what Protégé does) allows to focus more on the nature of objects than on defining their properties solely as binary relations that link them together (“roles” in DLs, “slots” in Protégé). Based on Aristotelian definitions by genus and differentia, concepts can be verbalized in a more human readable form than restrictions on roles. This type of definition is particularly useful for ontology extraction [13]. What is more, a terminological system, which is also an ontology authoring tool with logic-based formalisms and adheres to W3C standards, is extremely useful to domain experts and terminologists who do not have background in logic, but need to build their own machine-actionable and understandable domain terminologies. Tedi, a new ontology editing platform for terminologies of a given domain, was born out of the drive to respond to these needs. Formalized terminologies are essential for language processing tasks, for reasoning upon the data, for the creation of fully computable multilingual dictionaries, and for connecting object-based with text-based resources.

III. RELATED WORK

Relevant research on the state-of-the-art on representing the semantics of our data for the Semantic web points towards the following directions:

A. Ontologisation of non-ontological resources

A conceptual model of a domain is at the core of most knowledge based systems and language processing systems. The specific contribution of formal ontologies is the detailed, logical definition of the concepts and of the possible semantic relations between entities. Today one of the most prominent application of ontologies is the semantic indexing of content for resource discovery. This requires that the underlying data has rich and unambiguous semantics. The need for structuring the categories of the domain in a way that can be communicated without the risks of natural language ambiguity and polysemy has given rise to numerous efforts to use controlled languages and vocabularies. For a relatively recent state-of-the-art see [14]-[15].

This approach is similar to that of wielding the power of thesauri as a less powerful and less granular way to structure into a hierarchy the terms of a domain. Thesaurus structure concepts into monohierarchic trees or polyhierarchic lattices, ontologies structure them into semantically-rich directed graphs. The example of the ontologization of AGROVOC Thesaurus is a clear manifestation of the advantage of terminologically rich domain ontologies over other types of Knowledge Organization Systems (KOS) [16]. The current need to reengineer cultural heritage thesauri into ontologies is exemplified by Getty Vocabularies [17]-[18].

B. Building natural language interfaces for representing knowledge on the Semantic Web

Semantic Content Authoring and Linked Data authoring for user-friendly creation of content (manual or semi-automatic) on the web of data are rapidly emerging. Natural language interfaces support end users who are not computer experts. A range of capabilities such as the authoring of knowledge content, the retrieval of information from semantic repositories, and the generation of pattern for definitions in natural language make content management more intelligent through the injection of descriptive semantics in the process of content creation [19].

C. Building lexical models for the representation of lexical data on the Semantic Web

The primary mechanisms for the representation of lexical data on the Semantic Web has been the Lemon core model [20] (with extra modules for Syntax and Semantics, Decomposition, Variation and Translation, and Metadata [21]), further developed in the context of the W3C OntoLex community group into the new OntoLex-Lemon model [22].

D. Using existing ontology authoring environments

Before setting off, we considered using existing ontology editors. There exist different ontology editing tools, which support the creation and population of ontologies for the semantic web, but, to our knowledge, none which allows to
directly take into account the notion of ‘essential characteristic’ for defining domain concepts. In order to build our domain ontology, we used Tedi, a software which empowers domain experts to do their own ontological modelling. We decided against building our ontology directly in Protégé, even though Protégé is a feature-rich open-source platform for the construction of ontologies for the semantic web and is supported by a big user community. Protégé users have to familiarize themselves with defining classes (concepts) in terms of roles and role restrictions, which is hardly intuitive for those with no background in Logic. Granted, modelling in Protégé is a steep learning curve for non-computer scientists [23]. In contrast, Tedi supports the definition of formal ontologies by means of essential and descriptive characteristics, which are more intuitive to domain experts. For example, sewn is an essential characteristic of the garment exomis, whereas color is a descriptive one. Unlike descriptive characteristic, essential characteristic cannot be assigned a value. Its formalization requires a higher logic. Furthermore, the notion of ‘essential characteristic’ is a cornerstone for Conceptual Terminology in Specialized Languages. Conceptual Terminology distinguishes the definition of concepts (set of essential characteristics) from the description of objects (set of descriptive characteristics).

E. Ontologising cultural heritage

Last, in order to ontologise our terminology from the cultural heritage of ancient Greece we considered using relevant ISO standards, especially the ISO 1087-1 standard on vocabulary, theory and application of terminology, and the ISO 704 on principles and methods of terminology work. As already discussed these ISO standards are not operationalisable [11]. There is one ISO standard for the cultural heritage sector, which as will be shown below, our approach aims to extend. The vocabulary for the description of cultural objects was accepted as international standard ISO 21127 and is also known as CIDOC-CRM (Conseil International des Musées-Conceptual Reference Model) [24]-[25]. CIDOC-CRM does not specifically address the terminologies of the cultural heritage domain [26]. The same holds for other data models used in the cultural heritage and museum community, e.g., LIDO (Lightweight Information Describing Objects) [27] and EDM (Europeana Data Model) [28].

Due to the semantic richness and heterogeneity of cultural content and the distributed ways in which this content is created by domain experts, cultural heritage is a field where semantic technologies should become the standard technology to use. While archaeology and classical studies have spearheaded the use of digital tools, they have been quite slow in adopting W3C standards, mainly due to the belief that the type of humanistic inquiry pursued in these fields cannot or should not be standardized [29]. The theory and practice of ontological representation and modelling of archaeological, and more broadly, cultural heritage material, needs to be informed by the epistemic traditions of the disciplines involved [30]. Models that capture information independently of linguistic and cultural variation can standardize this diversity by adding a formal layer to the data. Knowledge, even tacit knowledge, needs to be expressed in a language, either natural or artificial. Models to cover both the conceptual and the terminological aspects of this knowledge are definitely going to multiply in the near future [31]–[33].

In the domain of ancient Greek cultural heritage, efforts are made to produce new domain-specific standards, such as the standard for digital editions of texts inscribed on a range of materials, including stone and papyrus (EpiDoc, Epigraphic Documents in TEI-XML, Text Encoding Initiative - eXtensible Markup Language) [34], and the Standards of Networking Ancient Prosopographies (SNAP) [35]. Moreover, geo-ontologies, such as Pelagios [36] and Google Ancient Places [37] link space as place to ancient time, while datasets of ancient artefacts, such as coins [38] and pottery [39], can now be published as LOD (Linked Open Data). The formalization of terminological systems in the domain, however, remains at a nascent stage.

IV. THE ONTTERMINOLOGY APPROACH

The need to make terminologies that are meant for human communication machine-processible according to international de facto and de jure standards motivated the first machine-readable trilingual terminology of ancient Greek dress (in English, French, and Greek) [40]. Our approach set out to build a formal domain ontology and make the resulting structured data shareable on the web of data. To achieve this means dealing with the ambiguity of natural language in defining the concepts of the domain. A degree of formalization/standardization was achieved, first, by clearly distinguishing between the concept level (i.e., the stable domain knowledge) and the term level (i.e., the natural language that is used to name the domain concepts); second, by putting them in relation (i.e., linking the terms in different languages to their denoted concepts). This leads to combining ontology and terminology into the new paradigm of ontotermology [41]. An additional objective was to create a tool that lowers the barrier for users not familiar with knowledge engineering, both at the technical level and at the level of the logical theory adopted. When exporting in OWL essential characteristics are translated into classes; essential characteristics belonging to the same axis of analysis, therefore mutually exclusive, are translated into disjoint classes. There are different ways of translating essential characteristics into OWL. The use of classes is one of them. It is also possible to simulate a second order logic in considering essential characteristics as individuals [42].

V. TEDI SOFTWARE PLATFORM

The Tedi software platform was developed in VisualWorks at the University Savoie Mont-Blanc [1]. It supports both term standardization and customization. Standardization of terminologies relies upon expert agreement on domain knowledge, which is necessary for collaboration and rapid sharing of information.

Tedi relies on a theory of concept inspired by the ISO standards on terminology. It is based on the notion of essential characteristic. The essential characteristics are
grouped into axes of analysis (sets of exclusive essential characteristics, e.g., a garment can be either wrapped or attached; either worn directly on the skin or as an overgarment; etc.). The set of axes of analysis constitutes an ‘orthogonal base’ for the meaning of the concepts. The logical properties of the system are verified at every step of ontology building.

Such modelling of domain knowledge can be very finely structured knowledge in order to eventually support two types of queries: by means of words, and by means of concepts. In order to clearly distinguish between the different types of knowledge on which Tedi relies, we use the following conventions: concepts are written between angle brackets “< >”, whereas essential characteristics (also called “differences” in Tedi) are written between slashes “/…/”, and terms written between quotes “...”. For example, the term “exomis” denotes a type of objects associated to the following set of characteristics: /for man/, /around body/, /more than one part/, /with sewing/, /without sleeves/, /attached/, /one attachment/, /knee-length/, /unpleated/, /undert/. 

A. Tedi Editors

Tedi’s rich architecture deploys two interconnected systems for the conceptual and linguistic dimensions. The concept editor allows to define essential characteristics, axes of analysis, attributes (descriptive characteristics), relations, and concepts. It also allows to update the ontology by inserting new concepts into the hierarchy. In order to help structure the system, Tedi automatically infers the possible generic concepts as well as the possible essential characteristics. The system’s in-built reasoner checks the compatibility of the essential characteristics in order to propose only those that are possible at a given moment. It also infers those that can be logically inferred and generates the formal definition of the concept, helping the expert to manage the combinatorial explosion (n axes of analysis made up of two exclusive essential characteristics potentially define 2^n concepts). If there is no concept corresponding to the set of essential characteristics denoted by a term, Tedi proposes to create a new concept and a new concept name based on the selected essential characteristics.

In the term editor, the user can: enter the terms in as many languages as needed, declare the status for each term (term status can be parameterized) and the part-of-speech for each term (choosing from: noun, verb, adjective, none), add contexts and notes. Tedi generates a pattern of definition for each term on the basis of the formal definition of the denoted concept. The system also calculates automatically the terminological equivalents across different languages, but also in a given language and for a given term the terminological synonyms, terminological hypernyms and hyponyms (two terms are terminological synonyms if and only if they denote the same concept).

B. Export Formats

Tedi enables domain experts to capture domain knowledge, to express it formally regardless of their background in formal languages, and to export it into different formats, which, of course, are not equivalent. At its present version (version 1.1) Tedi exports in CSV (Comma Separated Values), HTML (both static and dynamic), JSON, and RDF / OWL.

C. Use case: Conceptualizing Ancient Greek Garments

In the use case we present here the user needs to define the Greek dress multilingual ontotermology. Figure 1 shows a screenshot of the modelling of the garment termed “ἐξωμίς” in ancient Greek, “exomis” in English and “exomide” in French. Textual and iconographic evidence has shown that the “exomis” is a male unpleated and sleeveless garment that covers the body down to the knees and is attached at one point of attachment.

The ontotermology building process, centered on essential characteristics, consists in five interrelated non-linear iterative tasks that the expert should take for every concept defined in the system.

Task 1: Go to Tedi Term editor: enter the terms to be defined in the language(s) we need. These terms can be given directly by the experts or from NLP tools for candidate term extraction. Define their Status (choosing from the following drop-down list: preferred, alternative, tolerated, not recommended, obsolete), and their PoS (Part-of-Speech) (choosing from: noun, verb, adjective, none).

Task 2: In Tedi Concept editor: define the essential characteristics and the axes of analysis. These essential characteristics are found out by identifying differences between objects.

Task 3: In Tedi Term editor: link the term to the concept. Select all the essential characteristics that you want associated with the term. If there is no corresponding concept, Tedi proposes to create a new one, whose name is constructed from the chosen essential characteristics. The set of characteristics that have been selected is its formal definition, i.e., its definition in a formal language imbedded in Tedi. The axes of analysis, their dependencies and the compatibility of the essential characteristics are managed by Tedi. The system automatically checks the compatibility of the defined essential characteristics thus guiding the expert by proposing only those that are possible at a given moment.

Task 4: In Tedi Concept editor: update the ontology by inserting the newly created concept into the conceptual system, i.e., by linking it hierarchically (or associatively) with other concepts, supplementing its description by the addition of descriptive characteristics, if necessary. Where appropriate, new concepts can be introduced for the purposes of organizing the conceptual system without there being any terms that designate them in the given linguistic system. In order to help structure the system, Tedi automatically infers the possible generic concepts for a given concept, i.e., their intensional definitions, consisting in all their essential characteristics, are included in the intensional definition of the concept.

Task 5: In Tedi Term editor: complete the definition of terms in different languages. To this end, Tedi proposes ‘patterned’ definitions in natural language on the basis of the formal definition of the concept denoted by the term and its terminological hypernym. It remains for the expert to
reformulate them syntactically and put them in their final form.

D. Validation

Going back to the example, the concept denoted by the term “exomis” is defined by the characteristics: /for man/, /around body/, /more than one part/, /with sewing/, /without sleeves/, /attached/, /one attachment/, /knee-length/, /unpleated/, /under/. This set of features constitutes the formal definition of the concept. Tedi automatically infers that this concept counts <Garment around body> and <Garment for man> among its possible generic concepts. The concept name proposed by Tedi is a concatenation of these characteristics: <Garment for man around body more than one part with sewing without sleeves attached one attachment knee-length unpleated under >. The definition for the term ‘exomis’ in English is: “Short and non-pleated garment for man, usually worn around the body directly on the skin, this sleeveless garment consists of two sewn pieces of cloth, attached on the left shoulder leaving naked the right shoulder and part of the chest”. The Greek-English Lexicon, also known as LSJ (Liddell Scott Jones), which is the standard dictionary for scholarly use defines exomis as “tunic with one sleeve”. A mere comparison of the two definitions illustrates the usefulness of the essential characteristics approach.

VI. CONCLUSION & FUTURE WORK

To sum up, this paper presented a tool-assisted method for the ontologization of terminologies meant for human communication, so that they become interpretable also by machines. The approach and software presented here reflect the need for deeper semantics in the ontological part of a representation of reality, so that the represented part and the specialized language for human use can be more fully interpreted by machines. Tedi can be used to create multilingual terminological dictionaries of a domain containing definitions for terms in natural language, their canonical and inflected forms, and a wealth of related unstructured data in the form of notes, contexts of use, images, and videos. By combining ontology, terminology, and user-friendliness, Tedi software offers the possibility to enrich text-based data through semantic annotations. Ontotermologies can be exported into different interchange formats including JSON and OWL. An ontotermology mashup and server is currently under way.

ACKNOWLEDGMENT

We gratefully acknowledge the generous funding of the European Commission Marie Skłodowska-Curie (grant agreement no. 657898).

REFERENCES


Figure 1. Modelling the conceptual dimension for example domain concept denoted by term “exomis” in Tedi Concept Editor.