Instrumenting the UNWTO's Thesaurus on Tourism & Leisure Activities for seamless adoption in the Web of Data

Emanuela De Vita emy@crs4.it
CRS4 - Center for Advanced Studies, Research and Development in Sardinia Edificio 1, Parco Scientifico e Tecnologico, Località Piscina Manna 09010 Pula (CA) - Italy

Stefano Salvatore Fadda stefano@crs4.it
CRS4 - Center for Advanced Studies, Research and Development in Sardinia Edificio 1, Parco Scientifico e Tecnologico, Località Piscina Manna 09010 Pula (CA) - Italy

Alberto Farci farci@crs4.it
CRS4 - Center for Advanced Studies, Research and Development in Sardinia Edificio 1, Parco Scientifico e Tecnologico, Località Piscina Manna 09010 Pula (CA) - Italy

Gianluca Malagnini gianluca@crs4.it
CRS4 - Center for Advanced Studies, Research and Development in Sardinia Edificio 1, Parco Scientifico e Tecnologico, Località Piscina Manna 09010 Pula (CA) - Italy

Ivan Marcialis ivan@crs4.it
CRS4 - Center for Advanced Studies, Research and Development in Sardinia Edificio 1, Parco Scientifico e Tecnologico, Località Piscina Manna 09010 Pula (CA) - Italy

Keywords

Web of Data Ontology Thesaurus Tourism

Abstract

In this paper, we describe our conversion of the "Thesaurus on Tourism and Leisure Activities" [1] into SKOS. We propose well-formed and human-readable IRIs to identify every concept, and we connect most of these entities with the corresponding DBpedia entity. We also present an independent release of this thesaurus, it updates/removes the obsolete concepts, and supplements it with new ones. To manage these vocabularies, we developed a Web application that offers an API to query the model and the features to use and to share and, as a remarkable feature, allows users to improve the model by translating labels and descriptions in different languages and suggesting connections with different ontologies available in the Web of Data.

1. Introduction

After almost two hundred years since the mass tourism was born, [2] it arises as a huge network that reaches every corner of the earth and involves people from every social class, its influence thoroughly penetrates society, politics, culture and, above all, the economy: to design, produce and promote by using standards may become essential. In 2001, as response to this need, UNWTO [3] published the multilingual "Thesaurus on Tourism and Leisure Activities" [1] as a guide to tourism terminology, as well as standardization and normalization of a common indexation and research language at an international level. This thesaurus, hereafter referred to as TTLA, is an important starting point in the field of information technologies applied to travel, tourism, and hospitality industry; and several scholars refer to it to specify their models for the e-tourism

domain.

Although the UNWTO is still enhancing the thesaurus, the version published in 2001 remains the only one available and, after 15 years, brings to light some issues: a few concepts are outdated (EuropeanUnionCountries, USSR, Russia, CyberCafe, and others) and others are missing (InternetAccess, WiFi, HighSpeedConnections, and others); furthermore, thesaurus formats according to any standards provided by W3C for the Semantic Web obviously are missing too [2] .

In this paper we propose our contribution to re-engine and instrument the TTLA. We start presenting prior works. In the third section we describe our conversion of the TTLA to SKOS, the IRIs we propose to identify entities, and the upgrades of the original thesaurus. The last section describes the Web interface to manage the vocabularies, and the REST API to access them.

2. Related Works

As stated earlier, we based our work on standards of the Semantic Web that are the W3C's vision of the Web of linked data and particularly on SKOS. We also followed the W3C's best practices for publishing Linked Data [3] , specifically the fifth section " The Role of Good URIs for Linked Data " .

Concerning the SKOS conversion, we mainly referred to [4] ; the authors present a structured method to convert thesauri to SKOS, able to guarantee interoperability and completeness. Like this one, also our work is based on [5] , it presents guidelines and case studies for generating RDF encodings of existing thesauri with both standard (ISO 2788) and non-standard structure using the SKOS Core Schema. We also studied the TTLA SKOS versions as proposed by Mondeca [6] and by Junta Académica del Tesauro Turístico Argentino [7] but these works show some issues: both of them did not pay the proper attention to IRIs: those that Mondeca proposed are not well-formed and out of reach, and those that the Tesauro Turístico Argentino published do not follow any W3C recommendation [8] ; their SKOS versions are available only on their own Web sites, and there is no way to download, and therefore to efficiently use, the entire model; Mondeca does not provide any API to access its content; and the Tesauro Turístico Argentino provides labels only in Spanish.

Our Web application and REST API design draw liberally on [9] . We also followed the principles highlighted in SMIY [10] as well as in REST CookBook [11] .

3. Reengineering UNWTO's

In this chapter we present our conversion process of the TTLA thesaurus from its native format to SKOS with the aim of make the thesaurus available on the Web of Linked Data, providing a machine-readable format for sharing relevant information with a greater community.

3.1. Modeling the concepts IRIs

In order to provide a secure and permanent URL to every entity available in TTLA, as [12] suggests, we are using " Permanent Identifiers for the Web " [13] , a software project that offers a re-direction service for Web applications that deal with Linked Data and that often need to specify and use very stable URLs.

So, according to this service, the URIs have an *https* protocol, start with *w3id.org* , and are followed by the prefix *ttla* . Every concept is represented by its descriptor in Camel case form. Below are some examples:

https://w3id.org/ttla/CountriesAndCountryGroupings

https://w3id.org/ttla/Sports

3.2. Thesaurus conversion to SKOS

The conversion process was split up into three steps [4] : firstly we analysed the thesaurus, evaluating all the concepts and its relationships; then we applied the mapping, as shown in Table 1 , between the TTLA data items and the SKOS classes and properties; finally we performed the conversion following the defined

mapping.

For the analysis we used the TTLA in PDF available in [1] . The thesaurus is modelled with the ISO 5964 and 2788 standards and includes 8185 terms distributed in 3 language versions (French, English and Spanish). A structured alphabetical list, a hierarchical list, a list of descriptor groups (micro-thesaurus) and a permuted list are available for each language. The TTLA is organized into 20 semantic fields with a maximum for 5 levels of hierarchy. A semantic field represents all descriptors belonging to the same conceptual family. Each field carries the name of the concept (descriptor) that has the broadest meaning within the subject family. A descriptor is an expression that is not ambiguous, is standardized in its spelling. When the descriptor is the synonym of a non-descriptor the relationship is expressed by UF (Used For). A non-descriptor is a term that cannot be used for indexation because it is the synonym of a designated descriptor. The relationship is expressed by USE. In addition to this semantic equivalence, the TTLA relates the terms through three more kinds of relation: (a) hierarchy is indicated by the symbol BT for a Broader Term which has a larger meaning, or by the symbol NT for the Narrower Term which has a more specific meaning than its broader term; (b) multi-hierarchy, for descriptors that may simultaneously have several broader terms; and (c) association, that establishes a link between descriptors and it is indicated by RT that stands for Related Term. The TTLA also uses SN, for a Scope Note, it explains succinctly the semantic field assigned to a descriptor when it may be ambiguous. Most of the descriptors also have a serial number so that the classification is the same in all the languages of the Thesaurus.

The second step concerns to map every TTLA data item into SKOS schema. In the Table 1 below we summarize the mapping process [4] .

Data Item	Feature/function	Property/class
-	The whole TTLA	*skos:ConceptScheme*
Term A	Preferred term	*skos:Concept* with *skos:prefLabel* A and *skos:hiddenLabel* the Serial Number of A
Term A UF term B	Non-Preferred term	*skos:Concept* with *skos:prefLabel* A and *skos:altLabel* B
Term B USE term A	Non-Preferred term	*skos:Concept* with *skos:prefLabel* A and *skos:altLabel* B
Term C	A semantic field	*skos:Concept* with *skos:prefLabel* C TTLA *skos:hasTopConcept* C and C *skos:topConceptOf* TTLA
Term D BT term E	Broader term	D *skos:broader* E
Term E NT term D	Narrower term	E *skos:narrower* D
Term A SN " … "	Scope Note	A *skos:scopeNote* " … "
Term A has id 01.23	Serial number	A *skos:hiddenLabel* 01.23

Table 1. Mapping of TTLA Data Items to features and SKOS property/classes

Regarding the third step, we did not develop or use any specific software to automate the conversion process. Starting from the PDF version, we converted it into plain text and then, by means of a series of regular expressions, we replaced the identified patterns in the text document with the RDF elements and we used Protégé to manually verify every entity of the obtained SKOS document. Finally, we used the Web version of Skotify tool [14] to automatically improve, enrich and validate our TTLA conversion into a SKOS vocabulary.

3.3. A revised TTLA version

We provide two versions of TTLA. The first SKOS release keeps the original thesaurus unchanged and it is available at the URL:

https://w3id.org/ttla/v1.0/

The second release includes all the concepts previously present on UNWTO's work and improves the original model through new concepts and properties; this release is available at the URL:

https://w3id.org/ttla/v1.1/ (or) *https://w3id.org/ttla/*

Our revised TTLA version:

- introduces Italian labels for every concept;
- supplements the original thesaurus with new concepts (mainly countries, country groupings, novel technologies, and new travel services);
- updates the obsolete concepts or marks them as deprecated;
- adds a serial number to the terms devoid of;
- links, when possible, its concepts to the corresponding DBpedia resources, we use, according to SKOS specifications, skos:exactMatch, skos:closeMatch, and skos:relatedMatch.

4. The Web Application

Our Web applications are located at the URL *http://intuit.crs4.it/schemas* but, as mentioned earlier, in order to provide a secure and permanent URL, they are also available at *https://w3id.org/ttla* . These applications offer the following features:

- every SKOS Concept has an own Web page where all its characteristics are listed and explained;
- registered users can contribute improving the models by translating all the annotations in different languages;
- registered users can suggest connections with different ontologies available in the Web of Data.

Users suggestions are immediately and fully available just for the user who suggested them and she/he will be able to use her/his own contributions via the API right away. Only after these suggestions are considered and approved, they will become an integral part of the model.

4.1. The Rest API

In order to allow external applications to use our models, registered user can request an API key to access our basic Rest API [15] . In accord with [9] we thoroughly documented our API, providing examples, use cases, tips and guidelines on best practices, our API documentation is available at:

http://intuit.crs4.it/docs/api

We laid great importance to resource names and URL design, giving each resource a unique identifier, using plural nouns for collections, and using ids to access single resources.

Our API server supports HSTS (HTTP Strict Transport Security) to instruct users to never try insecure connections to it; the API base URLs use the following template:

*https://api.intuit.crs4.it/{ **graph** }/{ **ver** }/{ **type** }/{ **id** }*

Where:

- *{ **graph** }* is ttla;
- *{ **ver** }* is the TTLA version and is declared with the *v+number* , v1.0 refers to original thesaurus, v1.x to the upgraded models;

- *{ **type** }* is the list of the requested resource and can be *concepts* or *conceptSchemes* ;
- *{ **id** }* is the request resource identifier, it is optional.

As [9] recommends, the format of the resource's representation should be negotiated using the Accept header in client requests: actually we provide responses in JSON-LD; JSON; RDF/XML; Turtle; and Ntriples too.

5. Conclusion and Future Works

In this paper, we presented our work to utilize in the Web of Data the UNWTO's Thesaurus on Tourism and Leisure Activities. We studied the most accepted methods to convert a thesaurus from its native format to SKOS and tried to apply them to the TTLA. We pointed out and annotated the obsolete concepts, as some countries that not longer exist. When possible, setting the property *skos:exactMatch* , *skos:closeMatch* , and *skos:relatedMatch* with a DBpedia entity, we connected several concepts with the corresponding Wikipedia resource. We implemented a Web application to allow the scientific community to browse the thesaurus modelled in SKOS, and to permit registered users to improve the models translating labels and descriptions in different languages and connecting all the entities to different ontologies. We also published a Rest API, which makes the schemas available for third-party applications. We intend to carry on this work adding other connections with Dbpedia, and with different ontologies.

We would like to point out that this work is part of INTUIT project, funded mainly by Autonomous Region of Sardina [16] and developed thanks to the partnership between Space [17] , a company dedicated to enhance and promote the Italian cultural heritage, and CRS4 [18] , an interdisciplinary research centre. We intend to implement an exhaustive OWL 2 DL ontology to represent the whole field of tourism, we would want to propose a knowledge base focused on describing every content could suit in a tourist guide. This ontology will be further maintained and developed thanks by means of the same approach we described in this paper, we hope to involve a community for improving and test the ontology, we will offer Web applications and APIs for browsing and instrumenting our new ontology and enrich it with an exhaustive set of tourist contents related to our region, Sardinia.

References

1. UNWTO, 2001. Thesaurus on Tourism & Leisure Activities (Trilingual: English, French, Spanish). Retrieved from http://www.e-unwto.org/doi/book/10.18111/9789284404551
2. Berners-Lee T., Hendler J., and Lassila O., 2001. The Semantic Web, Scientific American, 29-37.
3. W3C, 2014. Best Practices for Publishing Linked Data. Retrieved from http://www.w3c.org/TR/ld-bp/
4. Van Assem M., et al., 2006. A Method to Convert Thesauri to SKOS. In Proceedings of the 3rd. European Semantic Web Conference (ESWC 2006). Springer, Heidelberg, 95-109.
5. Miles A. J., et al., 2004. Migrating thesauri to the semantic web - guidelines and case studies for generating RFD encodings of existing thesauri. Semantic Web Advanced Development for Europe project deliverable 8.
6. Mondeca Labs. Retrieved from http://labs.mondeca.com/skosReader/
7. Tesauro Turístico Argentino, Retrieved from http://www.tesauroturistico.gob.ar/tesauros/tesauro/index.php
8. W3C, 2008. Cool URIs for the Semantic Web. Retrieved from http://www.w3.org/TR/cooluris
9. Pintus A., and Pinna F., 2016. The Web API Design Guidelines for Happy Developers, Leanpub book. DOI: http://leanpub.com/thewebapinntux
10. A generalisation of the Linked Data publishing guideline, Retrieved from goo.gl/ZWAIBU
11. What is HATEOAS and why is it important for my REST API?, Retrieved from http://restcookbook.com/Basics/hateoas/

12. W3C, 2014. Best practices for publishing linked data. Retrieved from https://www.w3.org/TR/ld-bp/
13. Permanent Identifiers for the Web, Retrieved from https://w3id.org/
14. Suominen O., and Hyvönen E., 2012. Improving the quality of SKOS vocabularies with Skosify. In Proceedings of International Conference on Knowledge Engineering and Knowledge Management. Springer, Berlin Heidelberg, 383-397.
15. Fielding R. T., 2000. Representational State Transfer (REST). Chapter 5 of Architectural Styles and the Design of Network-based Software Architectures. PhD Thesis. University of California, Irvine. Retrieved from http://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm
16. Regione Autonoma della Sardegna, http://www.regione.sardegna.it/
17. Space S.p.A., http://www.spacespa.it/
18. Center for Advanced Studies, Research and Development in Sardinia, http://www.crs4.it/

Footnotes

1 We stipulated a provisional agreement with UNWTO to reuse and republish their copyrighted thesaurus, hereinafter called TTLA. [back]

2 Mass tourism is conventionally considered to begin in 1840s and Thomas Cook is seen as the inventor. [back]

3 Stands for United Nations World Tourism Organization: the United Nations agency for the promotion of responsible, sustainable and universally accessible tourism. [back]