

The Terminology Management Platform : A Tool for Creating Linked Open Data

Author: Eva Coudyzer, Royal Museums of Art and History, Belgium

Keywords: controlled vocabulary, thesaurus, semantic enrichment, thesaurus management, SKOS, Linked Open Data

Contact:

e.coudyzer@kmg-mrah.be

marie-veronique.leroi@culture.gouv.fr

Abstract

The Terminology Management Platform (TMP) is a web service developed in the cultural-heritage project AthenaPlus. Organizations can use the tool to publish and link controlled vocabularies in a semantic web-conform format.

The TMP proposes functionalities for vocabulary management, including: editing and creating vocabularies in an online collaborative environment; importing CSV and SKOS/RDF-formats; mapping between different vocabularies and linking to external resources; exporting and publishing in SKOS/RDF.

Mapping and enriching concepts in the TMP will result in more detailed and meaningful search results on the web, irrespective of language. This will valorise collections managed in cultural institutions on the web, as well as in online catalogues such as Europeana.

This short paper gives an overview of the current status and uses of TMP. The production version of the tool is planned for January 2015.

Controlled vocabularies, the semantic web and SKOS

A controlled vocabulary is described by ISO as a “prescribed list of terms, headings or codes, each representing a concept. Controlled vocabularies are designed for applications in which it is useful to identify each concept with one consistent label, for example when classifying documents, indexing them and/or searching them. Thesauri, subject headings schemes and name authority lists are examples of controlled vocabularies.”¹

The last decade controlled vocabularies have started playing a more prominent role in web information systems because of the manner in which they contain and represent knowledge. Vocabularies are believed to be powerful instruments in the realisation of the semantic web.

The semantic web is built on the principle of sharing and reusing data on the web to achieve better search results, irrespective of language. This can be done by automatically linking separate data on the web through the use of web links or URIs. In a controlled vocabulary each concept has an explicit hierarchical, associative or equivalent relation with the surrounding concepts. Preferred terms are accompanied by lead-in entries for synonyms or quasi-synonyms. This representation allows us to know that a “Berline” is a type of “[four-wheeled closed] carriage”, that “La Joconde” is a synonym of the “Mona Lisa” and that “Germany” is the English translation of “Deutschland”.

Using controlled vocabularies with the goal of giving meaning to the overwhelming amount of information on the web, will lead to greater visibility of individual collections and easier access to information for web visitors.

The World Wide Web Consortium (W3C) developed a format in RDF/XML which facilitates the encoding of information for sharing and interoperability across various computer applications. SKOS or *Simple Knowledge Organization System* was developed to “build a bridge between the world of knowledge organization systems – including thesauri,

¹ ISO 25964-2 :2013(E), p. 4



classifications, subject headings, taxonomies, and folksonomies – and the linked data community, with the goal of bringing benefits to both.”²

TMP: development history

The Linked Heritage project investigated in 2011 the various steps necessary for making data part of the semantic web³. It was concluded that an important drawback for organisations in the GLAM-sector for linking and publishing data on the web, was the poor technical expertise in mapping vocabularies, as well as the lack of an accessible tool to map and export vocabularies in the desired format.

A prototype of the TMP was created in the Linked Heritage-project and a production version will be released in January 2015 in the AthenaPlus project.

The TMP is developed and hosted by the research group Équipe Condillac of the University of Savoie in France, a technical partner in the AthenaPlus project. Other development partners include the Royal Museums of Art and History in Belgium and the Ministry of Culture and Communication of France.

TMP: Theoretical principles

The TMP is based on the notion of onto-terminology, a principle developed by Équipe de Condillac. It considers a terminology whose conceptual model is a formal ontology. The ontology refers to computational reasoning and a formal representation of concepts, whereas the terminology shows linguistic and multilingual variations. An “ontoterm” combines both the term definition and the concept definition. The “Onto-terminology engine” (OTe) is a tool which can explicitly represent concepts, terms and their definitions.

² Shiri A. (2012), p. 130-131

³ Cagnot et al. (2011)



The OTe can display hierarchical relations as is-a (BTG, NTG), part-of (BTP, NTP) and instance-of (proper name, BTI, NTI). The relations have logical properties: strict order (is-a), acyclic (is-a, part-of) and logical restraints (no specialisation or NTG, neither instantiation or NTI of individuals).

Specification sources for the development of the TMP are: onto-terminology for thesauri, ISO 1087-1, ISO 704, ISO 25964-1 and ISO 25964-2, as well as W3C for RDF, SKOS and OWL formats.

The TMP is fully RDF-based but makes use of the interchange formats SKOS and OWL.

TMP: Functionalities

Weblink to the TMP: athenaplus.thesaurus.condillac.org

Homepage

The homepage displays the logo of the TMP and of OTe, as well as the logos of Europeana, LinkedHeritage and AthenaPlus. The OTe logo also shows the TMP version number (“V2.0.12 2014-06-19”).

The TMP makes a difference between logged-in and not-logged in users. When you are not logged in, you have restricted rights and you can only view, scroll and search the vocabularies set as public by the proprietor. The TMP is only fully accessible when you have a username and a password. A login can be requested by sending an email to the administrator.

The homepage displays all uploaded or created controlled vocabularies set as public. Each vocabulary is described with a title, an author and a description. The vocabulary can be illustrated with images.

Once logged in you can see a change in the layout of the homepage. A distinction has been made between “My thesauri” and “Public thesauri”. “My thesauri” displays the vocabularies created or uploaded by yourself and the “Public thesauri” shows the thesauri from other providers set as public. There is a clear distinction between propriety thesauri and public thesauri.

Creating

Vocabularies can be created by entering title, author and description in a metadata box. The box also displays the creation and last modification date of the vocabulary. The author is automatically set as the person creating the vocabulary. Images for concepts can be added through the selection of a JPEG-image in your folders.

Vocabularies can also be imported from a SKOS or a CSV-file. The CSV-file can only be imported when conform a template, including clear attributions to SKOS-labels and properties.

Editing

The propriety vocabularies can be edited, deleted and exported in SKOS, Json and PDF. The editing functionalities include making alterations to concepts in the linguistic and the conceptual pane of the vocabulary, images, thesaurus structure, scope notes, translations etc.

Following the ISO-standards and the principles of OTe, the editing page is divided into two panes. The left pane – the linguistic dimension of the vocabulary – shows the thesaurus with the terms it contains in a specific language. The right pane displays the conceptual dimension of the concept: the hierarchical, associative and equivalent relations, the preferred and non-preferred terms and the links to external resources.

The language can be changed by selecting the ISO 639-1 language code from a drop-down list.

To define a hierarchical relation, both simple and more defined broader (BT) and narrower (NT) relations are available. A simple relation is a non-specific relation between a broader

and a narrower concept, e.g. the “Royal Museum of Art and History” is a NT of the concept “Museums”.

However, ISO25964-norms describe different types of hierarchical relations: generic (“sort of / type of”), whole-part (“is part of”) and instance (“is instance of”). These relations can also be specified in the TMP by choosing the correct BT (BTI, BTG or BTP) in the drop-down list. A specific relation would state that “Royal Museums of Art and History” is a BTI or “broader term instance” of “Museums”. Both options are available to the users.

Mapping

The TMP offers a mapping functionality which allows you to define equivalent relations between concepts from different vocabularies.

The link is set with ISO-standard symbols for equivalence relations: =EQ (exact equivalence) and ~EQ (inexact equivalence). =EQ and ~EQ correspond respectively to skos:exactMatch and skos:closeMatch.

Concepts can also be mapped by specifying the relation as broader (BM), narrower (NM) or related (RM).

The TMP: Accessibility and Use

The TMP is developed, managed and hosted by the University of Savoie, under the authority of the AthenaPlus project. The development of a thesaurus management tool for content partners is one of the objectives of WP4 “Terminology and Semantic Enrichment”.

In AthenaPlus, the TMP will be used to create and map semantically enriched controlled vocabularies. Concepts between vocabularies of the same domain can be mapped and linked to external resources, such as DBpedia. The mapped vocabularies can be published as linked



open data via a SPARQL endpoint and exported for reuse in other web applications or in local CMSs.

It is foreseen to use the TMP as an instrument for enriching digital content from AthenaPlus-providers in Europeana. This will be done by mapping local vocabularies to enriched reference vocabularies in the TMP. The reference vocabularies can be integrated in MINT - the webservice used in AthenaPlus for aggregating content - where they can be linked to the uploaded metadata. Hence, the enriched data from the TMP will be made available to visitors of Europeana.

Conclusion

The TMP is a webservice developed in the AthenaPlus project. It offers easy accessible functionalities to map and publish vocabularies as linked open data. Linking concepts from and between local and international standardized vocabularies will lead to greater visibility and optimized search results on the web and in online catalogues such as Europeana.

The TMP is based on the theoretical principle of onto-terminology and is compliant with ISO-standards. It includes following functionalities: editing, creating, deleting, searching, importing (SKOS, CSV), exporting (SKOS, Jason, PDF) and mapping.

In the near future more applications such as version management, online collaboration features, semi-automated mapping and a publishing environment will be included.

A production version of the TMP is planned for January 2015. Until the release of the production version, the TMP is a work in progress and liable to elaborate development or layout changes.

References

ATHENAPLUS (2013-2015)

www.athenaplus.eu

LINKED HERITAGE (2011-2013)

www.linkedheritage.eu

CAGNOT et al. 2011

Stéphane Cagnot, Johann Holland, Marie-Véronique Leroi, Michael Culture Aisbl, *Your Terminology as Part of the Semantic Web. Recommendations and Guidelines for Design and Management*, Paris 2011.

DAMAS, ROCHE 2013

Luc Damas, Christophe Roche, *First Release Terminology Management Platform (TMP)*, AthenaPlus D4.3, Le Bourget-du-Lac 2013

ISO 25964-2:2013(E)

Information and documentation – Thesauri and interoperability with other vocabularies – Part 2: Interoperability with other vocabularies, Geneva 2013

SHIRI 2012

Ali Shiri, *Powering Search. The Role of Thesauri in New Information Environments*, New Jersey 2012



