Submitted on: 23.01.2018

SKOSification of Trilingual Cultural Thesaurus (TCH) of National Library of Iran (NLI): A step in line with NLI's Linked Data strategy

Saeedeh Akbari-Daryan
Advisor to the Deputy Research, Planning and Technology of the National Library and Archives of Iran, Assistant Professor of the National Library and Archives of Iran
sakbaridaryan@gmail.com

Fariborz Khosravi
Deputy Research, Planning and Technology of the National Library and Archives of Iran. Assistant Professor of the National Library and Archives of Iran
fa.khosravi@gmail.com

Mahdi Ebrahimi
Software Architect of the National Library and Archives of Iran. Master of Arts in Computer Science
ebrahimi.mahdi@gmail.com

Hasan Bagheri
The project manager of Integrated Library Systems of National Library and Archives of Iran. The Bachelor of Arts in Computer Science
h_bagheri@hotmail.com

Abstract:

Considering recent developments in the fields of E-Culture and the role of thesauri as a part of the E-Culture demonstrator and the Linked Open Data (LOD) web, it seems to be quite obvious that Trilingual Cultural Thesaurus (TCH) of National Library of Iran (NLI) has to be made available on the web in a compatible for providing and sharing its relevant information with a wider community. This study discusses about the process of TCH SKOSification.

Keywords: Trilingual Cultural Thesaurus (TCH), National Library of Iran (NLI), SKOSification, SKOS, IranMARC, Content negotiation

Introduction

For the bibliographic and librarian world, Linked Data offers the technology and the social attention needed to publish and interlink metadata sets: the advantage is the access to all documents and resources indexed/classified/organized by means of the interlinked metadata sets (Morshed, Caracciolo, Johannsen & Keizer, 2011). The benefits of publishing library data as Linked Data have been recently summarized by the W3C Incubator Group on Library Linked Data. In particular, the following key benefits of Library Linked Data (LLD) have been identified: i) provides enhanced and more sophisticated navigation through information, ii) increases the visibility of cultural data, iii) supports integration of cultural information and digital objects into research documents and bibliographies, iv) offers a more durable and robust semantic model than metadata formats that rely on specific data structures, v) facilitates re-use across cultural heritage datasets, thus enriching the description of materials with information coming from outside the organization's local domain of expertise, and vi) allows developers and vendors to avoid being tied to library-specific data formats such as MARC (MAchine Readable Cataloguing) or Z39.50 (Vila-Suero, Villazón-Terrazas & Gómez-Pérez, 2013).

With the development of Simple Knowledge Organization System (SKOS), there is a standard way to represent knowledge organization systems using the Resource Description Framework (RDF). Due to the use of RDF, information can be used and re-used in a very interoperable way. As yet a lot of organizations and libraries are bringing their thesauri and vocabularies to the web in SKOS format (Zapilko and York Sure, 2009). SKOSification is a process of conversion of your thesaurus elements into a specific format. It means that the conversion is supported by rules, and that the result of such a process must be syntactically correct in regards with the format "grammar". Thus you have to check at the end if the SKOSified version of your thesaurus is correct or not.

"Our will to preserve and guarantee the existence, accessibility of documentary heritage to present and future generations" is in the Mission Statement of the National Library of Iran (NLI). Also "facilitation, acceleration and promotion of accessibility to the documentary heritage for audiences", and "development of the processing capabilities of organizing documentary heritage in order to make it more available" is one of the main strategic objects of NLI; Furthermore, "development of services with the highest level of quality, expected by various audience groups" is among the NLI strategic targets. Since users' satisfaction legitimatize the existence of the NLI, all efforts should move towards providing the highest and qualitative services (Strategic planning document, 2012).

It is important to know that the National Library of Iran as a trustee for managing the valuable documentary heritage is one of the few national libraries that provide organization of resources tools into national language. Some of these tools consist of *List of Persian Subject Headings*, *Trilingual Cultural Thesaurus* and *Persian Medical Thesaurus*.

Thesauri are possible building blocks of a web of Linked Data. As DBpedia for large data sets in general, specialized thesauri could be useful as interlinking hubs for professional communities – if they are available on the Linked Data web (Neubert, 2009).

Considering recent developments in the fields of E-Culture and the role of thesauri as a part of the E-Culture demonstrator and the Linked Open Data (LOD) web, it seems to be quite obvious that TCH of NLI has to be made available on the web in a compatible for providing and sharing its relevant information with a wider community. If, for example, a term in the

TCH is linked with a term in the UNESCO thesaurus, all documents indexed by the same term in the document repositories related to TCH and UNESCO are also potentially linked. Using appropriate applications, information queries can be submitted against both repositories, and data results presented (and processed) to the user in a unified way. For this reason, many thesauri are adopting the Linked Data approach to data publishing.

According to the above-mentioned, benefits of Linked Data (LD) and importance of thesauri in LD, as well as using SKOS in line with LD and on the other hand responding to call for papers from IFLA -global voice of the library and information profession- the National Library of Iran prioritized the TCH SKOSification process in its activities. After study phase, authors of the study provided the following phases to continue the rest of the project:

1. Developing mapping TCH/MARC to SKOS
2. Generating persistent URLs
3. Using Content-Negotiation Standard
4. Generating output inTCH/SKOS
5. Validating our SKOSification
6. Content validating the final migration of the TCH into SKOS format
7. Mapping TCH with well-known thesauri in the cultural domain

This study reports only phase1 to phase4 which could be completed until IFLA deadline for sending full paper. Final report of this project will be informed to the librarian interested in.

Related studies

Van Assem, Malaisé, Miles & Schreiber (2006) state to convert thesauri to RDF for use in Semantic Web applications and to ensure the quality and utility of the conversion a structured method is required. Moreover, if different thesauri are to be interoperable without complicated mappings, a standard schema for thesauri is required.

They present a method for conversion of thesauri to the SKOS RDF/OWL schema, which is a proposal for such a standard under development by W3Cs Semantic Web Best Practices Working Group. They apply the method to three thesauri: IPSV, GTAA and MeSH. With these case studies they evaluate their method and the applicability of SKOS for representing thesauri.

Summers, Isaac, Redding & Krech, (2008) describe a technique for converting Library of Congress Subject Headings MARCXML to Simple Knowledge Organization System (SKOS) RDF. The conversion and delivery of Library of Congress Subject Headings as SKOS has been valuable on a variety of levels. The experiment highlighted the areas where SKOS and semantic web technologies excel: the identification and interlinking of resources; the re-use and mix-ability of vocabularies like SKOS and Dublin Core; the ability to extend existing vocabularies where generalized vocabularies are lacking.

Neubert, J. (2009) in his/her article describes the conversion of a large economics thesaurus to RDF/SKOS, using the enhancement mechanisms of SKOS to dispose some nonstandard features of this thesaurus. The deployment, using RDFa pages, and the interlinking with other resources, namely a library catalogue and an experimental mapping to DBpedia are presented. For information retrieval support, a SPARQL query facility uses the data for building a thesaurus-backed terminology web service.

Zapilko, Schaible, , Mayr, &Mathiak (2013) in their article describe the conversion process of the Thesaurus for the Social Sciences (TheSoz) to SKOS. In order to create a semantically full representation of TheSoz in SKOS, extensions based on SKOS-XL had to be defined. These allow the modelling of special relations like compound equivalences and terms with ambiguities. Additionally, mappings to other datasets and the appliance of the TheSoz are presented.

The Archives of France (an agency of the Ministry of Culture) seized the opportunity of making available on the web the thesaurus for local archives to test Linked Data technologies. Since January 2011, a website makes it possible to browse the thesaurus, to download RDF/XML documents but also to search via SPARQL which is the query language for RDF structured data (see: http: //data. culture. fr/thesaurus/). This project was also an opportunity to align data with other controlled vocabularies and resources (thesaurus RAMEAU and DBpedia) and to implement a solution for persistent identifiers of concepts of the thesaurus. Sibille-de Grimoüard, (2014) presents this open government data initiative, aiming not only to establish an interoperability framework between archival data, but also to consider archival data with other cultural data.

Trilingual Cultural Thesaurus (TCH) of National Library of Iran (NLI)

Persian Cultural Thesaurus as the first Iranian thesaurus written in Persian in 1995 by the Islamic Revolution Cultural Documentation Organisation (IRCDO) affiliated to the Ministry of Culture and Islamic Guidance. After the merger of IRCDO and NLAI in 1999, second edition of Persian Cultural Thesaurus was published in 2001.

The development of Trilingual Cultural Thesaurus (TCH) Persian –English- Arabic has begun by NLI since 1997. TCH was published in 2006 and the second edition was published in 2013. TCH serves widely as a crucial instrument for indexing the content of non-book materials in Iran, e. g. Islamic Republic of Iran Broadcasting. Also, the educational centres use it in order to teaching of indexing course. This thesaurus is owned and maintained and used for indexing especially for audio-visual materials by NLI. Records of TCH as an authority files are kept IranMARC authorities format. Also, In NLI's database has stored over the 800000 bibliographic records of audio-visual materials covering 80 different format (e. g. pictures, slides, DVD video, tapes, stamps, etc.), documents, Three Dimensional Resources, oral history and other resources which utilized TCH. TCH covers all topics and sub-disciplines of culture in 18 domain (e. g. Education, Arts and Culture, History, etc.). It should be noted that the language of Russian and Tajik script is being added to the TCH. Findings of conceptual analysis of the thesaurus are summarized below:
TCH contains 11124 preferred (authorized) terms, 3232 non-preferred (unauthorized) terms, 7645 narrower terms, 3629 related terms, 14248 English terms (authorized and unauthorized) and 14239 Arabic terms (authorized and unauthorized). Additionally a classification hierarchy is provided and each thesaurus term is dedicated to one or more classification terms.

1. Developing mapping TCH/MARC to SKOS

IranMARC National Committee was established in 1998 at the National Library of Iran. This committee has developed IranMARC based on UNIMARC and by using -9- or 9-- in fields, subfield 9 and indicator 9 for national use with permission under the Permanent UNIMARC Committee (PUC).
This committee has implemented IranMARC bibliographic format, authorities and holdings so far. Now there are about two million bibliographic records for more than 70 type of

materials and about 700, 000 authority records including Persian Subject Headings, Personal Names, Corporate Body Names and TCH Records, etc. in NLI database. There are also about four million holding records stored based on IranMARC.

The IranMARC Authorities format distinguishes between authorized (2XX) and non-authorized (4XX) headings. Similarly the SKOS vocabulary provides two properties, skos: prefLabel and skos: altLabel, that allow a concept to be associated with both preferred and alternate natural language labels. In general, this allows authorized and non-authorized TCH to be mapped directly to skos: prefLabel and skos: altLabel properties .

SKOS has been designed for use in a multi-lingual environment. SKOS users are encouraged to use language tags to identify the language of particular label:
ex: animalsrdf: typeskos: Concept;
skos: prefLabel "animals"@en;
skos: prefLabel "animaux"@fr(Summers, Isaac, Redding & Krech, 2008).

Permanent UNIMARC Committee (PUC) recommends two options for multi–lingual controlled vocabularies: 1) Fields of 7XX. 2) Repeatable fields of 2XX (Willer,2009)

IranMARC National Committee has used the second option (repeatable fields of 2XX) for TCH's English and Arabic authorized terms. So the authors of this study should were seeking for an indicator through data of field 250 to distinguish Persian, Arabic and English terms from each other.

UNIMARC uses subfield 7 to specify Script of Cataloguing and Script of the Base Access Point. In IranMARC subfield 7 is used for Script of the Base Access Point. This subfield is a two-character alphabetic code specifies the script of the base access point when the identical access point appears in the record in a different script.

In TCH/IranMARC records, Subfield 7 is empty for terms in Persian language. For terms in English language the value “*ba”* and for Arabic terms the value *fa* is dedicated to subfield 7. So if *250$7=*” “ then *skos: prefLabellang=”fa”* , *250$7=ba* then *skos: prefLabellang=”en”* and if *250$7=fa* then *skos: prefLabellang=”ar”*.

IranMARC uses the 5XX fields to link an authorized heading to other related authorized headings. SKOS provides a rich set of semantic relationships between conceptual resources, including: skos: related, skos: broader, skos: narrower. The values “g” , “h” and “9” are used to display the relationship information (broader term, narrower term and related term) in TCH/IranMARC records respectively. It means that if *550$5=g* then *SKOS: broader*, if *550*$5=h then *SKOS: narrower* and if *550 $5=9* then *SKOS: related*. Thus the semantic relationships present in TCH/IranMARC are easily translated into TCH/SKOS.

TCH notes placed on fields 320 and 330 of UNIMARC. These fields are General Explanatory Reference Note and General Scope Note respectively.

The SKOS vocabulary also includes documentation properties which can be used to represent TCH/SKOS: skos: definition, skos: scopeNote. These properties are easily converted from TCH/MARC to TCH/SKOS.

As previously mentioned, TCH covers all topics and sub-disciplines of culture in 18 domains. These domains place on Iran/MARC in field 152 and subfield 9 that all mapping to skos: Concept. Finally TCH/SKOS Mapping results are detailed as you see in Table 1.

TABLE 1-TCH/SKOS Mapping

RDF property	**IranMARC/TCH**
skos: Concept	152$9
skos: prefLabellang=”fa”	250$a if $7=” ”
SKOS: prefLabellang=”en”	250$a if $7=”ba”
SKOS: prefLabellang=”ar”	250$a if $7=”fa”
SKOS: altLabel	450$a
SKOS: scopeNote	330$a
SKOS: definition	320$a
SKOS: broader	550$a if $5=g
SKOS: narrower	550$a if $5=h
SKOS: related	550$a if $5=9

1. Generating persistent URLs

As SKOS data are expressed as RDF triples. The concepts may be subject or object and related via a SKOS property which would be the predicate. As RDF triples, SKOS concepts can be identified using URIs. Although the SKOS data model does not require the use of persistent identifiers, the exploitation of SKOS concepts in Linked Opened Data necessitates their use.

Since every IranMARC Authority record supplied by NLI contains a “Record identifier” in the 001 field, it makes a good candidate for the identification of SKOS concepts. Record identifiers are guaranteed to be unique. SKOS requires that URIs are used to identify instances of skos: Concept. Semantic Web technology—as specified by RDF— and Linked Data practices also encourage the use of HTTP URLs to identify resources, so that resource representations can easily be obtained (Summers, Isaac, Redding &Krech, 2008). Of course data of 001 field that named NLIID are not URLs, so the data of 001 field are normalized and then incorporated into a URL using the following template http://opac.nlai.ir/opac-prod/authority/{NLIID}#concept. The authors of the articles preferred using the 001 field in concept identifiers, because terms of TCH are in constant flux, while the 001 for a record remains relatively constant. Persistence also allows metadata descriptions that incorporate TCH/SKOS concepts to remain unchanged, since they reference the concept via a persistent URL. The use of hash URIs for SKOS concept simplifies the web server implementation; since the server isn’t required to redirect using a 303 See Other HTTP status code, when the URI for the concept is requested.

It is important to know that the authors of the articles preferred using Semantic URLs to represent TCH/SKOS concepts on the Web. Semantic URLs, also sometimes referred to as clean URLs, RESTful URLs, user-friendly URLs, or search engine-friendly URLs, are Uniform Resource Locators (URLs) intended to improve the usability and accessibility of a website or web service by being immediately and intuitively meaningful to non-expert users. Such URL schemes tend to reflect the conceptual structure of a collection of information and decouple the user interface from a server's internal representation of information. Other reasons for using clean URLs include search engine optimization (SEO), conforming to the representational state transfer (REST) style of software architecture, and ensuring that individual web resources remain consistently at the same URL. This makes the World Wide Web a more stable and useful system, and allows more durable and reliable bookmarking of web resources. (Wikipedia, 2016)

2. Using Content-Negotiation Standard

The authors chose to deliver multiple representations of TCH/SKOS concepts on the Web using a technique called content-negotiation. When deciding what content to deliver to an HTTPclient, a web server can examine the Accept header sent by the client, to determine the preferable representation of the resource to send The TCH/SKOS delivery application currently returns the following representations: rdf/xml, application/html representations.
http://opac.nlai.ir/opac-prod/authority/{NLIID} to represent as html format
http://opac.nlai.ir/opac-prod/authority/{NLIID}.rdf to represent as rdf/xml format

3. Generating output inTCH/SKOS

The process of programming conversion and delivery TCH/SKOS, should be done with Java Script because the database of IranMARC is based on Java Script. It was done using the following open-source libraries:

- MARC4J[1]: The goal of MARC4J is to provide an easy to use Application Programming Interface (API) for working with MARC and MARCXML in Java
- Apache Jena[2]: A Java API for RDF

[1] https: //github. com/marc4j/marc4j

- Marc4j is used to pars and read TCH/MARCwhich provides an object-oriented, streaming interface to MARCXML records, and Apache Jena is used to create RDF format of records. Apache Jena is a free and open source Java framework for building Semantic Web and Linked Data applications. A relational database is used to store both raw data and relation of each record.

Conclusion

TCH SKOSification is valuable for NLA in several aspects. This is:

- Good start to optimal use of linked data.
- First Iranian multilingual thesaurus placed on SKOS.
- Useful experience for more effective participation in Semantic Web.
- An experience provides a suitable context for converting other Persian Thesauri and Subject Headings.
- A suitable pattern for other countries that use UNIMARC-based Data Storage Systems for their thesauri. They can accelerate converting data to SKOS

References

- Morshed, A. , Caracciolo, C. , Johannsen, G. , & Keizer, J. (2011). Thesaurus alignment for linked data publishing. Accessed may 25, 2016http: //eprints. rclis. org/21106/
- Zapilko, Benjamin, and York Sure. "Converting the TheSoz to SKOS. " *GESIS Report* (2009). Accessed may 25, 2016 http: //www. gesis. org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2009/TechnicalReport_09_07. pdf
- Zapilko, B. , Schaible, J. , Mayr, P. , &Mathiak, B. (2013). TheSoz: A SKOS representation of the thesaurus for the social sciences. Semantic Web, 4(3), 257-263.
- Strategic planning document (2012). National library & Archives of I. R. Iran . Deputy of research planning & technology.
- Vila-Suero, D. , Villazón-Terrazas, B. , & Gómez-Pérez, A. (2013). datos. bne. es: A library linked dataset. Semantic Web, 4(3), 307-313.
- Summers, E. , Isaac, A. , Redding, C. , &Krech, D. (2008). LCSH, SKOS and linked data. UniversitätsverlagGöttingen, 25.
- Sibille-de Grimoüard, C. (2014). The Thesaurus for French Local Archives and the Semantic Web. Procedia-Social and Behavioral Sciences, 147, 206-212.
- Neubert, J. (2009). Bringing the" Thesaurus for Economics" on to the Web of Linked Data. *LDOW, 25964*.
- Willer, M. (Ed.) (2009). UNIMARC Manual: Authorities Format (Vol. 38). KG Saur VerlagGmbh& Company, 2009.
- Van Assem, M. , Malaisé, V. , Miles, A. , & Schreiber, G. (2006). A method to convert thesauri to SKOS (pp. 95-109). Springer Berlin Heidelberg.
- Wekipedia (2016). Semantic URL. Accessed may 25, 2016https: //en. wikipedia. org/wiki/Semantic_URL

[2]https: //jena. apache. org