PROPOSAL TO REPRESENT THE UNESCO THESAURUS FOR THE SEMANTIC WEB APPLYING ISO-25964

Juan-Antonio Pastor-Sánchez (1)

(1) University of Murcia, Faculty of Communication and Documentation, pastor@um.es

Abstract

This paper shows how has been applied ISO-25964 standard to represent the UNESCO thesaurus through semantic web technologies. Based on the works done by the UNESKOS project, has been analyzed the joint use of SKOS and ISO-THES ontologies to represent thesauri according to the data model of the ISO-25964 standard. The result has been an RDF dataset, accessible as Linked Open Data, using SKOS and ISO-THES with the properties of a vocabulary developed in the context of UNESKOS project. The conclusions point, among other aspects, a review of SKOS and adoption of appropriate technologies that facilitate the development of future works and research lines focused on the alignment of vocabularies.

Keywords: UNESCO Thesaurus; SKOS; Semantic Web; ISO-25964; ISO-THES; UNESKOS

1 Introduction

Actually, SKOS is the most used ontology to represent interoperables Knowledge Organization Systems in the Semantic Web (Pastor-Sánchez, et al., 2012). SKOS was developed in a first stage into the SWAD-EUROPE project between 2002 and 2004. Finally was adopted as W3C recommendation in 2009 (W3C, 2009). Along its development the main purpose was the representation of a wide variety of controlled vocabularies into a simple way, i.e. Thesauri.

For many years the normative references for the elaboration of thesauri was the ISO-5964:1985 and ISO-2788:1986 standards. Some minor revisions was done regarding structural and terminological elements, mainly with the ANSI/NISO Z-39-19:2005. First proposals of deeply changes about the concept of Thesaurus was done by the British Standard BS-8723. As a result of those works, currently Thesauri are seen as a tool for the terminological control and the information retrieval that complains interoperability mechanism between different types of knowledge organization systems.

The Use Cases and Requirement document, redacted during the SKOS creation process, considered the ISO-5964/2788 standards. However, ISO-25964 was published after SKOS has came as an standard "de facto" in the context of the Semantic Web for thesaurus publishing. This implies difficulties for the deployment of thesaurus conforms to ISO-25964 in Semantic Web.

Since the publication of ISO-25964 the need to adapt SKOS to the new standard was identified. For this, a working group are developing the ISO-THES ontology that complements SKOS and complains the properties of ISO-25964 for representing any kind of Thesaurus (Isaac and De Smedt, 2015).

The UNESKOS project exists since 2012 with the aim to offer SKOS versions of both, the Proposed International Standard Nomenclature for Fields of Science and Technology and the UNESCO Thesaurus. The project applies the Linked Open Data principles (Pastor-Sánchez, et al., 2013). This work is about the update of UNESKOS project for the UNESCO Thesaurus and its publication as a SKOS dataset, available according to Linked Open Data principles and conforms to ISO-25964 standard. First, the methodology followed is described. Next section exposes the obtained results to apply ISO-THES and a comparative respect the previous SKOS version is done. Finally the paper points the conclusions and several future work lines.

2 Methodology

In order to update the SKOS representation of the UNESCO thesaurus have been analyzed the elements of the ISO-THES ontology. Next, it has been defined the mapping between such elements and the structure of the UNESCO Thesaurus. The two SKOS versions of the Thesaurus have been compared to check the viability of the new modeling.

Since a new modeling is proposed, it was also desirable using web scraping to obtain a new digital version of the on-line UNESCO Thesaurus and further processing

using regular expressions (Schrenk 2012 p.49-75) to obtain normalized text files for each language.

These files were processed to obtain a representation of the thesaurus using SKOS and ISO-THES. Additionally, the algorithm implemented for this purpose have distinguished between existing concepts in the previous version and those newly created to maintain the URI of existing concepts and collections.

After this process, the RDF/Turtle and RDF/XML serialization were obtained and used to update the thesaurus data in the triplestore. It should be noted that the RDF data stored in the triplestore are used to generate the serialization in different formats and for a SPARQL Endpoint as shown in Figure 1.

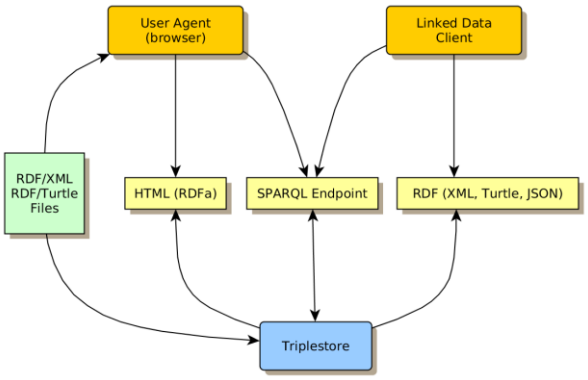

Figure 1: Functional diagram for data access of UNESKOS project. Source: PASTOR-SÁNCHEZ et al., 2013.

After this update was necessary to adapt the script to generate the HTML view to display the new elements used for modeling the thesaurus, specifically those corresponding to ISO-THES ontology and the ad-hoc properties developed for the UNESKOS project.

3 ISO-THES Ontology: Confluence of ISO-25964 and SKOS

One of the most significant features of SKOS is the definition of two levels of representation: the conceptual level and the lexical level. The class **skos:Concept** allows to define concepts to which terms are associated in different languages by assigning preferred labels (**skos:prefLabel**), alternative labels (**skos:altLabel**) and hidden labels (**skos:hiddenLabel**).

Using the terminology from the thesaurus standards, before ISO-25964, preferred labels correspond with descriptors and alternative labels with non-descriptors.

The hierarchical and associative semantic relationships are defined in the conceptual level. That is, the concepts, not the labels, are the elements connected by semantic relationships. These relationships identify generic (**skos:broader**), narrow (**skos:narrower**) and related (**skos:related**) concepts. This approach allows an efficient maintenance and application in indexing tasks of thesauri, since with SKOS a change in the concepts labeling does not imply any change in the structure of the semantic relations.

Additionally SKOS provides other constructs to define grouping structures, such as concept schemes (**skos:ConceptScheme**) and collections (**skos:Collection**) with its corresponding properties.

The definition and maintenance of the lexical level of the thesaurus with the new standard ISO-25964 can be relatively complex, due to the existence of equivalence relationships between terms (**USE/USE+** and **UF/UF+**) that complement the relationships between terms and concepts (**isNonPreferredlabelFor**, **isPreferredlabelFor**).

These relationships allows to define the equivalence between a non-preferent composed term (**SplitNonPreferredTerm**) with several preferred terms (**PreferredTerm**) using a composed equivalence relationship (**CompoundEquivalence**). SKOS has no elements to define these equivalences, not even with SKOs-XL (SKOS eXtension for Labels).

On the other hand, the part two of ISO-25964 standard deals the interoperability of thesauri with other controlled vocabularies. A typology of equivalences between concepts of different vocabularies is defined:

- Exact equivalence: both concepts, in source and target vocabularies, represent the same idea and both have identical scope. Example: “Mad cow disease” and “Bovine spongiform encephalopathy”.
- Inexact equivalence: concepts that have overlapping scopes, or a small difference of connotation or context of use. Example: “seats” and “chairs”.
- Partial equivalence: concepts with a close meaning although one is slightly broader than the other. Example: "Aircraft" and "Air-planes".
- Hierarchical mappings: this mapping is used when one concept is clearly broader than the other. Example: “Netherlands” and “Holland”.
- Associative mappings: this mapping is used when it is not possible define any equivalence or hierarchical mapping between two concepts, but are semantically related. Example: “Photographs” and “Photographers”.

Equivalence can be defined between two concepts of different thesauri (simple equivalence) or when a concept of a thesaurus is represented by combining several concepts in another thesaurus (compound equivalence).

In this last case the equivalence can be established with the intersection or the union of the meaning of the thesaurus target concepts. Equivalence relations are not

explicitly reflected in the data model of the standard, but can be derived from hierarchical and associative relationships.

SKOS allows interconnection of vocabularies, but only for simple equivalences. It does not include the new composite equivalence of the new ISO standard. The property **skos:exactMatch** provide the mechanism to define the ISO exact equivalence, while **skos:closeMatch** is used for inexact or partial equivalences. The properties **skos:broadMatch**, **skos:narrowMatch** and **skos:relatedMatch** are used for the hierachical and associative mapping relationships.

Therefore, there are many similarities between the new ISO-25964 standard and SKOS:

1. Both offer models of representation that can be applied to the development of information retrieval applications as shown in Figure 2 in Appendix.
2. The thesaurus is structured into two levels: conceptual and terminological (lexical).
3. ISO-25964 and SKOS have higher structures for concepts grouping. It is possible to represent thesaurus, concept groups and concept arrays with the ISO model. SKOS provides skos:ConceptScheme and skos:Collection to represent such structures.
4. Annotation elements are almost identical in both (see table 1).

Anyway we must consider that the new ISO standard was published when SKOS had been widely deployed to publish controlled vocabularies in the Semantic Web. An ontology has been developed to facilitate the interoperability, which reuses elements of SKOS to represent thesaurus in the Semantic Web context according to ISO-25964 (Isaac; de Smedt, 2015).

It also defines a number of elements for expressing the semantic richness of the ISO model. This paper is not intended to present in detail the ISO-THES ontology, however it is essential to point out some of the general characteristics:

- The SKOS and SKOS-XL classes are reused to represent concepts, concept schemes and collections. Two new grouping structures are defined: "Thesaurus Array" (**iso-thes:ThesaurusArray**) and "Concept Group" (**iso-thes:ConceptGroup**). The properties to link concepts with concept schemes, arrays, collections and concept group are also re-used and/or defined.
- The object properties to represent the semantic relationships between concepts are reused. However, It is necessary to define sub-properties from **skos:broader**, **skos:narrower** and **skos:related** in order to achieve the potential of the ISO data model and represent different types of hierarchical and associative relationships.
- The ISO standard defines generic, partitive and instances hierarchical relationships with a more precise semantics than **skos:broader** and **skos:narrower**. In any case it is possible to define the ISO hierarchical relationships as sub-properties of SKOS hierarchical relationships.
- The SKOS labeling and annotation properties are totally reused. In addition, SKOS-XL is essential to define the elements to represent the composite equivalence of ISO-25964.

Annotation element	**ISO 25964**	**SKOS**
General notes	Note	skos:note
Scope notes	ScopeNote	skos:scopeNote
History notes	HistoryNote	skos:historyNote
Release notes	EditorialNote	skos:editorialNote
Definitions	Definition	skos:definition
Examples	-	skos:example
Change notes	-	skos:changeNote
Custom notes	CustomNote	-

Table 1: Comparison between ISO-25964 and SKOS annotation elements

4. Proposal of representation of UNESCO Thesaurus

The biggest problems encountered by the team of the UNESKOS project were associated with the modeling of the UNESCO Thesaurus with SKOS. More specifically, the main difficulty lay in the representation of the knowledge areas and micro-thesauri, since SKOS does not provide any native element for a direct representation of these structures.

At the beginning of the project four options were studied, discarding the alternative that included the use of ISO-THES, since this ontology was in the initial stage of its design. It was decided to represent the whole thesaurus as a concept scheme. Areas and micro-thesauri was represented as collections. Areas was connected to the thesaurus using **skos:inScheme** and micro-thesauri with their respective areas and concepts with **skos:member**.

For example, assuming that <S> represents the UNESCO Thesaurus, <S1> represents the knowledge area "Education", <S105> represents the micro-

thesaurus “Educational sciences and environment” and <C02240> represents the concept “Learning” the representation was done as follows:

```
<S> rdf:type skos:ConceptScheme .

<S1> rdf:type skos:Collection ;
  skos:inScheme <S> ;
  skos:member <S105> .

<S105> rdf:type skos:Collection ;
  skos:member <C02240> .

<C02240> rdf:type skos:Concept
```

However, the first modeling version of the UNESCO Thesaurus with SKOS had some deficiencies from the point of view of its publication as Linked Open Data. These deficiencies limit the capacity to find information from the links between the elements of the vocabulary. More specifically: SKOS has no property to define a link from a concept to the collection to which it belongs.

In the case of the UNESCO Thesaurus this means that it is not possible to determine the micro-thesaurus to which a concept belong, considering only the explicit RDF data about the concept. The first solution designed in the UNESKOS project obtains this information from a SPARQL query. However, we understand that an approach based purely on Linked Open Data requires an explicit statement for the relationship between a concept and the collection to which belongs.

Furthermore, another deficiency was identified regarding the definition of the top concepts of a micro-thesaurus. This aspect is of great importance because the top concepts are the starting point for exploring the hierarchical structure. In the current modeling, this information is obtained from the intersection of two sets: the first one formed by all concepts of thesaurus defined as top concepts and the second one formed by all concepts that belong to certain collection. Applying exclusively SKOS it is impossible to determine the top concepts of a micro-thesaurus using the RDF data of a collection.

As in the previous case, this problem was solved using SPARQL queries. The query to obtain the primary concepts for micro-thesaurus "Education Policy" can be seen en the example code 1 in Appendix.

However, a software that discover SKOS concepts based only on links between resources can not get this information.

These limitations affect the representation of the dataset, the process of obtaining a complete description using the statements of the concerned resource only and the RDFa markup (W3C, 2015) of the HTML serialization of the thesaurus.

To solve these problems three solutions was analyzed:

Option A) Inclusion of a reverse semantic markup in HTML view. This is done using the **rev** attribute. This RDFa attribute is used to specify inverse relationships between two resources from the object to the subject of the RDF statements (see example code 2 in Appendix). The **rev** attribute reserverses semantics of subject and object.

Option B) Application of Concise Bounded Application: This technique include along with the description of a resource (Stickler, 2005), other statements in which the described resource is an object (see example code 3 in Appendix). It can be seen a description through the RDF statements in which the concept in the subject. A statement in which this resource is the object to denote the membership to a micro-thesaurus is also included. This option is controversial, because the community of developers and the dataset publishers do not have a single line of thought in this regard [1].

Option C) Application of ISO-THES ontology and definition of an Ad-hoc vocabulary with complementary properties (Pastor-Sánchez, 2015). Using both would express the relationships between elements of the thesaurus in both directions.

Options A and B were discarded, since it was considered that option C supposed a single solution with a more general approach. Thus, the problems of representation, resource description and semantic mark-up of HTML view were solved simultaneously.

In order to apply ISO-THES, the following modelling criteria were established:

- The whole thesaurus is represented by a concept scheme (**skos: ConceptScheme**).
- Areas and micro-thesauri are defined as groups of concepts (iso-thes: ConceptGroup).
- The areas are linked to the thesaurus using the property **iso-thes:microThesaurusOf**.
- Micro-thesaurus are linked with areas by **iso-thes:supergroup** and vice versa with **iso-thes:subgroup**.
- To indicate the concepts that belong to a micro-thesaurus has been applied the property **skos:member**.
- The property **uneskos:memberOf** have been defined as inverse of **skos:member** to represent the micro-thesaurus to which a concept belongs.
- The property **uneskos:mainConcept** have been defined to represent the top concepts of a micro-thesaurus.
- The property **uneskos:mainConceptOf** have been defined as inverse of uneskos:mainConcept.

The definition of properties **uneskos:mainConcept**, **uneskos:mainConcept** and **uneskos:memberOf** reflects the intention of facilitating the discovery of resources directly from the HTML view because RDFa is used for semantic markup. Without these properties it is impossible to determine:

- The top concepts of an **iso-thes:ConceptGroup** class resource. The properties **skos:hasTopConcept** and **skos:topConceptOf** can represent the top concepts of concept schemes only, while **iso-thes:ConceptGroup** is a subclass of **skos:Collection**.
- The micro-thesaurus to which a concept belongs. It is important to distinguish between this and the possibility of indicating the concepts contained into a micro-thesaurus, for which **skos:member** is used. The definition of the relationship **uneskos:memberOf**, lets to define a link from the concept to its micro-thesaurus, while skos:member define the link from the micro-thesaurus to the concept.

Consequently, using the new modeling, the representation shown at the beginning of this section would be as follows:

```
<S> rdf:type skos:ConceptScheme .

<C02240> rdf:type skos:Concept ;
  uneskos:memberOf <S105> ;
  uneskos:mainConceptOf <S105> .

<S1> rdf:type iso-thes:ConceptGroup ;
  iso-thes:microThesaurusOf <S> ;
  iso-thes:subGroup <S105> .

<S105> rdf:type iso-thes:ConceptGroup ;
  iso-thes:superGroup <S1> ;
  skos:member <C02240> ;
  uneskos:mainConcept <C02240> .
```

Thesaurus <S> is represented as a Concept Scheme. Knowledge Area <S1> and micro-thesaurus <S105> are represented as concept group using the class **iso-thes:ConceptGroup**.

Area <S1> is linked with the Thesaurus <S> using the property **iso-thes:microThesaurusOf** and the **iso-thes:subGroup** defines the inverse relationship.

Merbership of micro-thesaurus <S105> to the knowledge area <S1> is defined through the **iso-thes:superGroup**, using **iso-thes:subGroup** for the inverse link.

The property **uneskos:memberOf** establishes the membership of the concept <C02240> with the micro-thesaurus <S105>.

Finally properties **uneskos:mainConceptOf** and **uneskos:mainConcept** define <C02240> as top concept of the micro-thesaurus <S105>.

5 Conclusions

After creating the new representation of the UNESCO thesaurus, it has been observed that there has been no substantial changes from the previous version (the thesaurus still having 4408 concepts). The administrators of the UNESCO Thesaurus contacted with the team of UNESKOS project to identify and resolve inconsistencies found during the first modeling version of RDF data set. Since January 2013, the HTML version of the UNESCO thesaurus at UNESKOS project has received more than 80,000 visits.

During the modeling work with ISO-THES, the UNESKOS team could check the suitability of this ontology respect to the data model of the ISO 25964 standard. There is no doubt in this regard. However, in the author's opinion SKOS would require a review to incorporate into its core the inverse relationship of skos:member and the possibility of identifying top concepts into the collections. This would make it unnecessary to use uneskos properties: uneskos:memberOf, uneskos:Mainconcept and uneskos:mainConceptOf defined for semantics expressiveness in the context of the Linked Open Data publication.

Currently are being evaluated the adoption of Apache Jena Fuseki to support the triplestore of the UNESKOS project. The ARC2 framework, which runs on PHP and MySQL, has been quite reliable and efficient, but its evolution has stopped and the incorporation of SPARQL 1.1 is not provided, which limits the development of UNESKOS project.

A search engine based on Apache Solr is under development. This could solves many of the problems posed by the use of spelling and lexical variants of the terms entered in the query. Apache Solr in aligning vocabularies is being applied, with results that seems interesting, but only based on the lexical structure of the vocabularies. A future research line may be the vocabularies alignment through the structure of semantic relationships.

Notes

(1) For more information about the implications of Concise Bounded Description and different approaches: http://linkeddatabook.com/editions/1.0/#htoc43

References

ANSI/NISO (2005). Z39.19:2005 Guidelines for the construction, format and management of monolingual controlled vocabularies. Bethesda, MD: NISO Press, 2005.

British Standards (2005). BS 8723-2:2005 Structured vocabularies for information retrieval. Guide. Thesauri. London: British Standards Institution, 2005.

British Standards (2007). BS 8723-4:2007 Structured Vocabularies for information retrieval. Guide. Interoperability between vocabularies. London: British Standards, 2007.

ISO (1985). ISO 5964:1985 Guidelines for the establishment and development of multilingual thesauri. Geneva: International Organization for Standardization. 1985.

ISO. 1986. ISO 2788:1986 Guidelines for the establishment and development of monolingual thesauri. Geneva: International Organization for Standardization, 1986.

ISO (2011). ISO 25964-1:2011 Thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval. Geneva: International Organization for Standarization, 2011.

ISO (2013). ISO/DIS 25964-1:2013 Thesauri and interoperability with other vocabularies. Part 2: Interoperability with other vocabularies. Geneva: International Organization for Standardization, 2013.

ISO TC46/SC9/WG; Isaac, Antoine (2013). Correspondence between ISO 25964 and SKOS/SKOS-XL Models. http://www.niso.org/schemas/iso25964/correspondencesSKOS.

Isaac, Antoine; De Smedt, Johan (2015). [ISO-THES] // http://pub.tenforce.com/schemas/iso25964/skos-thes (2015-10-01).

Pastor-Sánchez, Juan Antonio (2015). UNESKOS Vocabulary // http://skos.um.es/TR/uneskos/

Pastor-Sánchez, Juan Antonio; Martínez-Méndez, Francisco Javier; Rodríguez-Muñoz, José Vicente (2012). Aplicación de SKOS para la interoperabilidad de vocabularios controlados en el entorno de linked open data. // El profesional de la información, 21:3 (mayo-junio 2012) 245-253.

Pastor-Sánchez, Juan-Antonio; Martínez-Méndez, Francisco-Javier; López-Carreño, Rosana; Rodríguez-Muñoz, José Vicente (2013). UNESKOS: publicación como Linked Open Data de la Nomenclatura Internacional de Ciencia y Tecnología y del Tesauro UNESCO. // I Congresso ISKO Espanha e Portugal/XI Congresso ISKO Espanha // http://eprints.rclis.org/24272 (2015-10-01).

Schrenk, Michael (2012). Webbots, spiders, and screen scrapers: A guide to developing Internet agents with PHP/CURL. San Francisco: No Starch Press, 2012.

Stickler, Patrick (2005). CBD - Concise Bounded Description. W3C Member Submission 3 June 2005 // http://www.w3.org/Submission/2005/SUBM-CBD-20050603 (2015-10-01).

W3C (2009). SKOS simple knowledge organization system reference. W3C Recommendation 18 August 2009. Miles, Alistair; Bechhofer, Sean (eds.) // http://www.w3.org/TR/skos-reference (2015-10-01).

W3C (2015). RDFa Core 1.1 Third Edition: Syntax and processing rules for embedding RDF through attributes. W3C Recommendation 17 March 2015. Adida, Ben; Birbeck, Mark; McCarron, Shane; Herman, Ivan (eds) // http://www.w3.org/TR/2015/REC-rdfa-core-20150317 (2015-10-01)

Received: 2015-11-03. Accepted: 2015-12-15.

Appendix

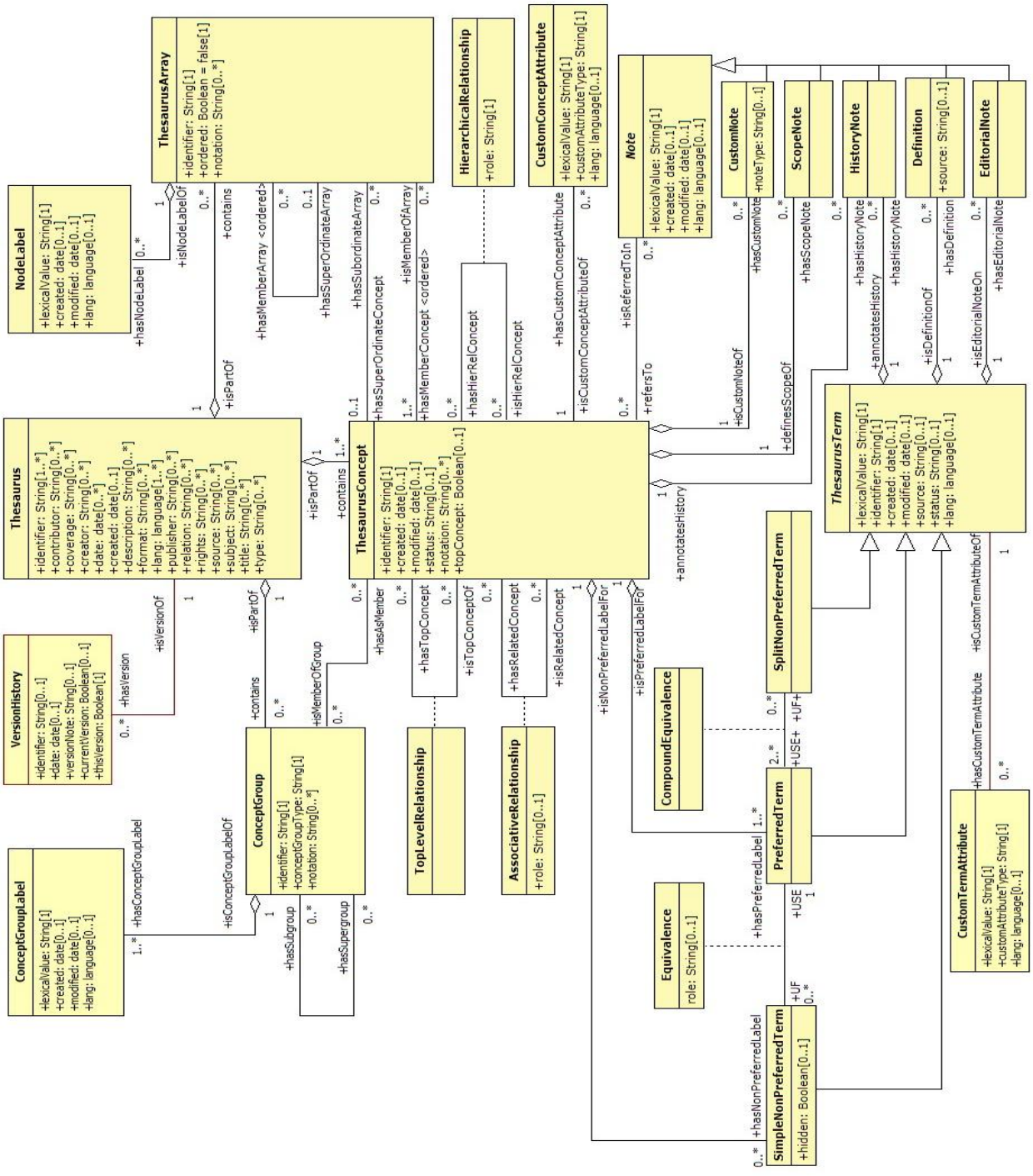

Figure 2: ISO-25964 UML data model. Source: <http://www.niso.org/schemas/iso25964/Model_2011-06-02.jpg>

```
SELECT ?concept {
        FROM <http://skos.um.es/unescothes/> {
                ?concept skos:topConceptOf <http://skos.um.es/unescothes/CS001>.
                <http://skos.um.es/unescothes/COL110> skos:member ?concept .
        }
}
```

Example code 1:

```
<div id="node" about="http://skos.um.es/unescothes/C01184" typeof="skos:Concept">
  <h2><span property="skos:prefLabel" xml:lang="en">Economics of education</span>
  <small>(http://skos.um.es/unescothes/C01184)</small></h2>
  ...
  <h3>Microtesauri</h3>
  <ul rev="skos:member">
    <li resource="http://skos.um.es/unescothes/COL115">
      <strong>MT</strong> <a href="http://skos.um.es/unescothes/COL115/html">
      1.15 Educational planning</a>
    </li>
  </ul>
  ...
</div>
```

Example code 2:

```
unescothes:C01184 rdf:type skos:Concept ;
  skos:prefLabel "Экономика образования"@ru, "Economía de la educación"@es,
    "Économie de l'éducation"@fr, "Economics of education"@en ;
  skos:inScheme unescothes:CS000 ;
  skos:topConceptOf unescothes:CS000> ;
  skos:narrower unescothes:C01280, unescothes:C01256, unescothes:C01232,
    unescothes:C01224, unescothes:C01221, unescothes:C01197, unescothes:C01196 ;
  skos:related unescothes:C01284, unescothes:C01180 ;
  skos:scopeNote "Экономические методы, применяемые в образовательных системах."@ru,
    "Techniques de l'économie appliquées au système éducatif."@fr,
    "Técnicas de la economía aplicadas al sistema educativo."@es .

unescothes:COL115 rdf:type skos:Collection;
  skos:member unescothes:C01184 .
```

Example code 3: