Modelling Knowledge Organization Systems and Structures

A discussion in the context of conceptual and data models

Maja Žumer, University of Ljubljana, Slovenia

Marcia Lei Zeng, Kent State University, USA

Abstract

In the last few decades, knowledge organization systems (KOS), especially thesauri, classification schemes and lists of subject headings, have largely followed or conformed with the established data models defined by standards, recommendations or best practices. This long list contains some widely used models, such as ISO5964 Part 1, ISO2788, Z39.19, BS 5723 and BS 6723 (Dextre Clarke, 2008), IFLA Principles Underlying Subject Heading Languages (SHLs), and MARC 21 Format for Classification Data.

The FRSAD (Functional Requirements for Subject Authority Data) conceptual model is the third member of the FRBR family, developed under the auspices of IFLA. The report was approved in 2010 and will be published in 2011. FRSAD is a general conceptual model that focuses on the subject relationship and therefore provides a theoretical framework for all KOS and their data models. In addition, it also assists in the assessment of the potential for international sharing and (re)use of subject authority data both within the library sector and beyond.

In this paper FRSAD is compared to SKOS and SKOS XL as data models (with implementation examples).

Background: the core of the FRSAD conceptual model

Within the IFLA FRBR framework, the core of the FRSAD model contains three entities (*work, thema, and nomen*) and two basic relationships: "work has as subject thema" and "thema has appellation nomen" (Figure 1) (Functional Requirements for Subject Authority Data, A Conceptual Model , 2010, Sections 3.4 Thema and 3.5 Nomen). In the model, *thema* is defined as "any entity used as a subject of a *work."* *Nomen* is "any sign or sequence of signs (alphanumeric characters, symbols, sound, etc.) by which a *thema* is known, referred to or addressed as."

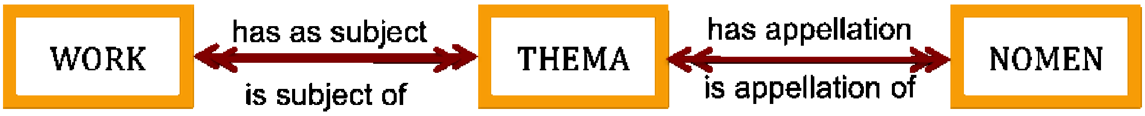


Figure 1: The basic FRSAD model.

It is important to note that the *thema* entity class is not restricted to actual subjects of *works* within a particular collection, but rather, anything that is or has the potential of being or becoming a subject. This generality not only enables the development of different knowledge organization systems (KOS) and tools, but allows for different implementations according to particular circumstances and needs as well. In regards to the granularity of a *thema*, it should be understood that *thema* can be the totality of what a particular *work* is about and/or any of the more atomic aspects of that totality. We have occasionally noticed a misconception that *thema* only represents the total body, set or sum of ideas contained in a work. Putting aside the argument that in reality such a total sum of ideas contained in a work may not be objectively determined, it should be obvious from the definition that an instance of *thema* of a particular *work* can be anything the cataloguer (or rather, the future user) considers as part of the 'has as subject' relationship.

Focus of this article

It is the purpose of this article to discuss the importance of a conceptual model and its implementation in data models for concepts and for other entities. This paper's focus is on the second part of the model. The 'has appellation/is appellation of' relationship is a new relationship defined by FRSAD. The notion had been introduced by FRAD, Functional Requirements for Authority Data, a conceptual model developed before the FRSAD model, in the form of several appellation entities ("name", "controlled access point", and "identifier"). FRSAD therefore combines and generalizes these entities and relationships. The FRSAD 'has appellation /is appellation of' relationship' is a many-to-many relationship in general. Any *thema* will have more *nomens* (e.g. in different languages, in different knowledge organization systems). In a natural language, a *nomen* may be an appellation of more than one *thema*: 'crane' in English is used both for an animal and engineering equipment. However, in controlled vocabularies, the situation of one *nomen* being used for more than one *thema* is avoided and each *nomen* may only be the appellation of one *thema*. To achieve this, qualifiers or other methods of disambiguation are used. Figure 2 shows the appellation relationship in a controlled vocabulary.

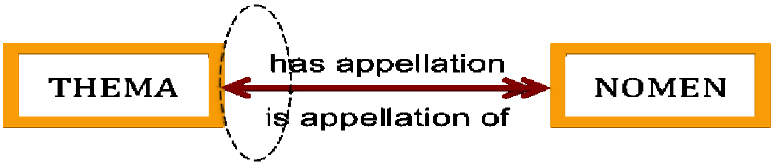


Figure 2: The 'has appellation' relationship between *thema* and *nomen* in a controlled vocabulary

The implications of the FRSAD model can be discussed in the context of the structures of subject authority systems or types of knowledge organization systems (KOS). The basic elements in any KOS structure can be analyzed from the FRSAD perspective regarding its:

1. Entities
   - Thema
   - Nomen
2. Relationships:
   - Thema – to – nomen
   - Thema – to – thema
   - Nomen – to – nomen
3. Attributes
   - Attributes of thema
   - Attributes of nomen

One of the fundamental notions of the *thema-nomen* model for subject authority data is to separate *themas* from what they are known as, referred to, or addressed as. This fundamental notion is consistent with KOS structures. Emphasizing this aspect will help the common understanding of KOS principles and the development of any knowledge organization structures.

The FRSAD conceptual model and its implementation in data models for subject authority data

This paper will use the data models specified by the W3C SKOS (Simple Knowledge Organization System) documents, including its optional extension for labels (SKOS-XL). SKOS is "a common data model for knowledge organization systems such as thesauri, classification schemes, subject heading systems and taxonomies. Using SKOS, a knowledge organization system can be expressed as machine-readable data. It can then be exchanged between computer applications and published in a machine-readable format in the Web." (SKOS Simple Knowledge Organization System Reference, 2009)

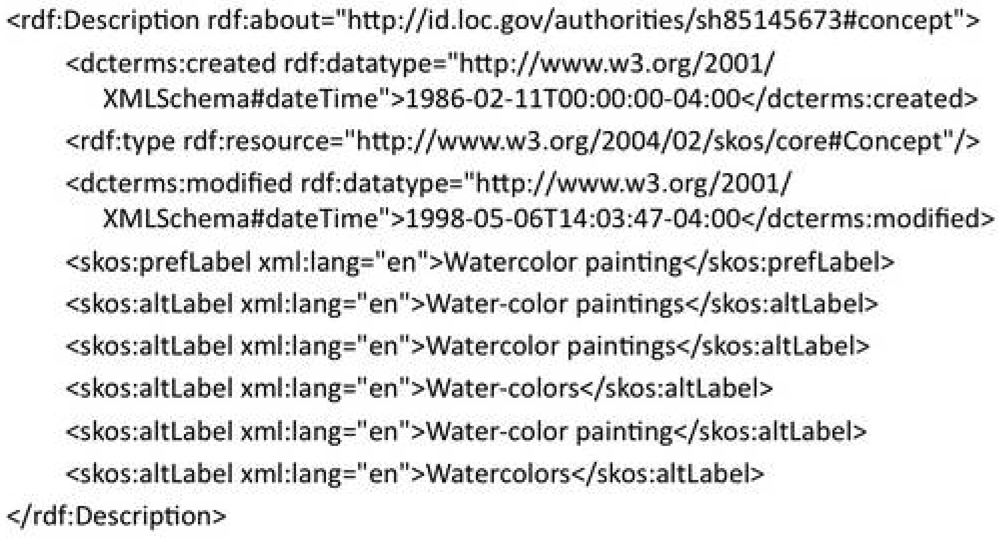
1. SKOS and the thema-nomen relationship model.

The FRSAD conceptualization of the relationship between a concept (thema) and the representation(s) of the concept (nomen(s)) echoes the SKOS Simple Knowledge Organization System Core data model back in 2005 (SKOS Core Vocabulary Specification, 2005), when the FRSAR (Functional Requirements for Subject Authority Records) Working Group was started.

The SKOS Core model clearly emphasized a concept-centric view of vocabulary, where primitive objects are not labels; rather, they are concepts represented by labels. The root of the model can be found in the thesaurus standards developed before SKOS Core, but such an emphasis was not clearly stated or modeled due to the mix of relationships of concepts (e.g., Broader Term (BT), Narrower Term (NT), and Related Term (RT)) and between the concept and its labels (Use and Used For (UF)). The use of the word 'term' in semantic relationships of broader and narrower concepts reflects such a mixed representation. In the SKOS Core model, labels (preferred, non-preferred, and hidden) are affiliates of a concept while the semantic relationships exist among concepts. "Mirroring the fundamental categories of relations that are used in vocabularies such as thesauri [ISO2788], SKOS supplies three standard properties" (SKOS Simple Knowledge Organization System Primer, 2009) for semantic relationships: skos:broader and skos:narrower for hierarchical links and skos:related for associative (non-hierarchical) links. These convey the same relationships between themas defined in the FRSAD model.

The KOS vocabularies that have implemented this SKOS core (or no extension for labels) model can be found in those already published as Linked Data, such as the Library of Congress Subject Headings (LCSH) at http://id.loc.gov/authorities/ (See example of "watercolor paintings' as human-readable format at http://id.loc.gov/authorities/sh85145673 and as RDF/XML format at http://id.loc.gov/authorities/sh85145673.rdf).
Let's take a look at the concept and label handled in this example. (Semantic relationships between concepts, such as broader, narrower, related, and the scope note of the concept are not copied here. See original full RDF/XML record at: http://id.loc.gov/authorities/sh85145673.rdf)

<http://id.loc.gov/authorities/sh85145673#concept>



```
<rdf:Description rdf:about="http://id.loc.gov/authorities/sh85145673#concept">
    <dcterms:created rdf:datatype="http://www.w3.org/2001/
      XMLSchema#dateTime">1986-02-11T00:00:00-04:00</dcterms:created>
    <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
    <dcterms:modified rdf:datatype="http://www.w3.org/2001/
      XMLSchema#dateTime">1998-05-06T14:03:47-04:00</dcterms:modified>
    <skos:prefLabel xml:lang="en">Watercolor painting</skos:prefLabel>
    <skos:altLabel xml:lang="en">Water-color paintings</skos:altLabel>
    <skos:altLabel xml:lang="en">Watercolor paintings</skos:altLabel>
    <skos:altLabel xml:lang="en">Water-colors</skos:altLabel>
    <skos:altLabel xml:lang="en">Water-color painting</skos:altLabel>
    <skos:altLabel xml:lang="en">Watercolors</skos:altLabel>
</rdf:Description>
```

Figure 3. Extracted statements for a concept that has a label in English, "Watercolor painting". Source: LCSH RDF/XML record http://id.loc.gov/authorities/sh85145673.rdf

In this entry, the concept has a unique identifier, <http://id.loc.gov/authorities/sh85145673#concept>. Its preferred label has the value "Watercolor painting". It has several alternative labels two among them being "Watercolors" and "Water-color paintings".

These extracted statements can be illustrated by the figure below:

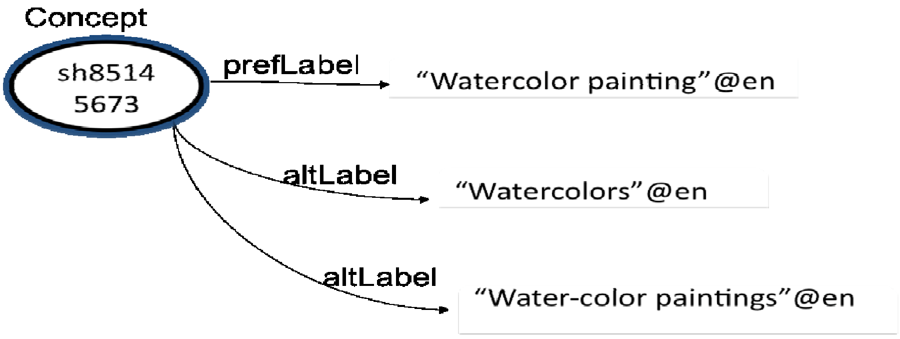


Figure 4. Simplified illustration of a subject heading entry, where labels are attributes of the concept.

This data model treats appellations of a thema as its surrogates. To explain in entity-relationship model terminology, this means:

1. The Concept entity has attributes such as "preferred label" and "alternative label".
2. The appellations (strings of characters) are attribute values of the concept attribute.
3. Labels have no attributes of their own.
4. There is no relationship between labels.

These characteristics are different from the data model based on SKOS + SKOS-XL, which is discussed in the next section.

2. SKOS eXtension for Labels (SKOS-XL) and relationships of nomens.

The FRSAD model's addition of an entity to the original proposed FRBR model: nomen, is also quite significant. This enables the treatment of the so-called "label" of concept to become an entity itself, which allows one to define attributes of this entity as well as relationships between instances of a nomen.

This aspect is paralleled by the newer version of the SKOS, which supplements an eXtension for Labels (SKOS-XL) (SKOS Simple Knowledge Organization System Reference, 2009, Appendix B) specification in 2009. This version defines an extension, providing additional support for identifying, describing and linking lexical entities. To align with the SKOS 2009 specification, *thema* corresponds to skos:Concept class and *nomen* corresponds to skosxl:Label, as illustrated below:

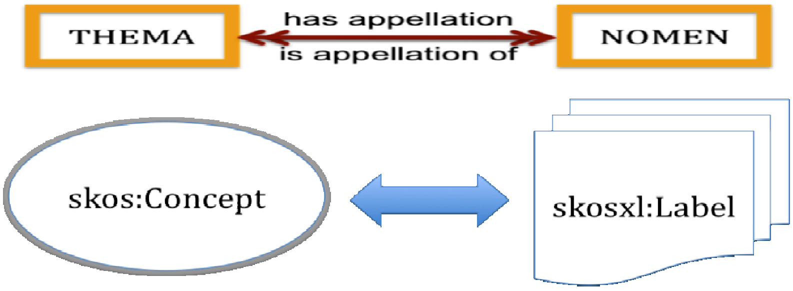
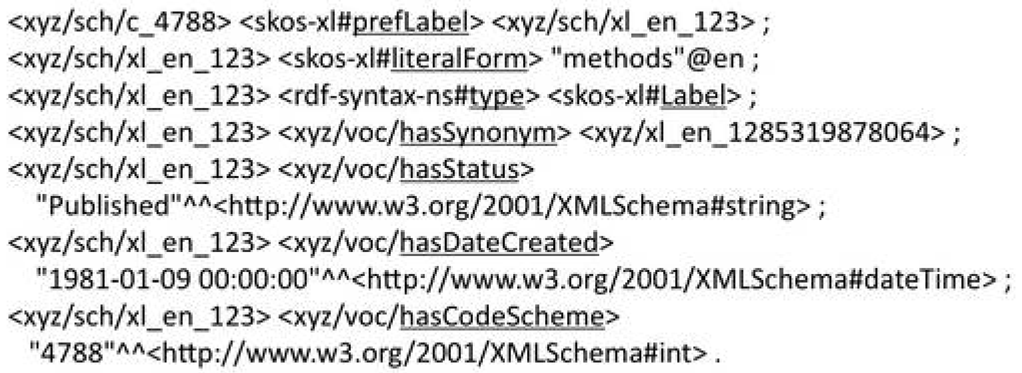


Figure 5. Aligning FRSAD model with SKOS + SKOS-XL data model.
The capitalization of *Concept* and *Label* indicates that they are classes.

The FRSAD model defines *nomen* as an entity. *Nomens* have attributes as well as relations between or among themselves while representing the same *thema*. This can be illustrated with a situation when a preferred label of a concept in a concept scheme has various literal forms, synonyms, status of release, and administrative data. FRSAD has provided a few common possible relationships and is flexible: implementation- or domain-specific relationships can be added. With the availability of SKOS-XL, such situations can be handled appropriately.

Taking an example of a multilingual thesaurus AGROVOC that is available as Linked Data, the concept and the label classes clearly stand as separate entities. Each preferred label has literal form, synonym, status, creation date, and other administrative information. The following figure is a simplified example created to explain the data model. Short names for the namespaces are used. For example, a concept's URI in XYZ's concept scheme, "http://xyz/schemename/xl_en_123", is shortened as "xyz/sch/xl_en_123"; a property defined by XYZ's vocabulary "http://xyz/voc/hasStatus" is shortened as "xyz/voc/hasStatus" and the property name is underlined; and "http://www.w3.org/2008/05/skos-xl#literalForm" is shortened as "skos-xl#literalForm". As indicated by the codes in the identifiers, "c" represent a concept (e.g.,'c_4788' is the ID for a concept instance) and "xl" represents a label where its language is also indicated by the language code, e.g., "xl_en_123" is the ID for English preferred label of concept 'c_4788'.



```
<xyz/sch/c_4788> <skos-xl#prefLabel> <xyz/sch/xl_en_123> ;
<xyz/sch/xl_en_123> <skos-xl#literalForm> "methods"@en ;
<xyz/sch/xl_en_123> <rdf-syntax-ns#type> <skos-xl#Label> ;
<xyz/sch/xl_en_123> <xyz/voc/hasSynonym> <xyz/xl_en_1285319878064> ;
<xyz/sch/xl_en_123> <xyz/voc/hasStatus>
   "Published"^^<http://www.w3.org/2001/XMLSchema#string> ;
<xyz/sch/xl_en_123> <xyz/voc/hasDateCreated>
   "1981-01-09 00:00:00"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
<xyz/sch/xl_en_123> <xyz/voc/hasCodeScheme>
  "4788"^^<http://www.w3.org/2001/XMLSchema#int> .
```

Figure 6. Simplified statements, based on AGROVOC example,
for a concept that has a label in English, “methods”

When multiple languages are involved, this same data model will be further extended to account for each language label(s). This structure is illustrated by Figure 7.

- ‘c_4788’ is the URI for a concept instance.
- ‘xl_en_123’ is the URI for English preferred label of concept ‘c_4788’.
- ‘xl_de_789’ is the URI for German preferred label of the same concept.
- ‘xl_en_345’ is the URI for a synonym of the preferred label ‘xl_en_123’.

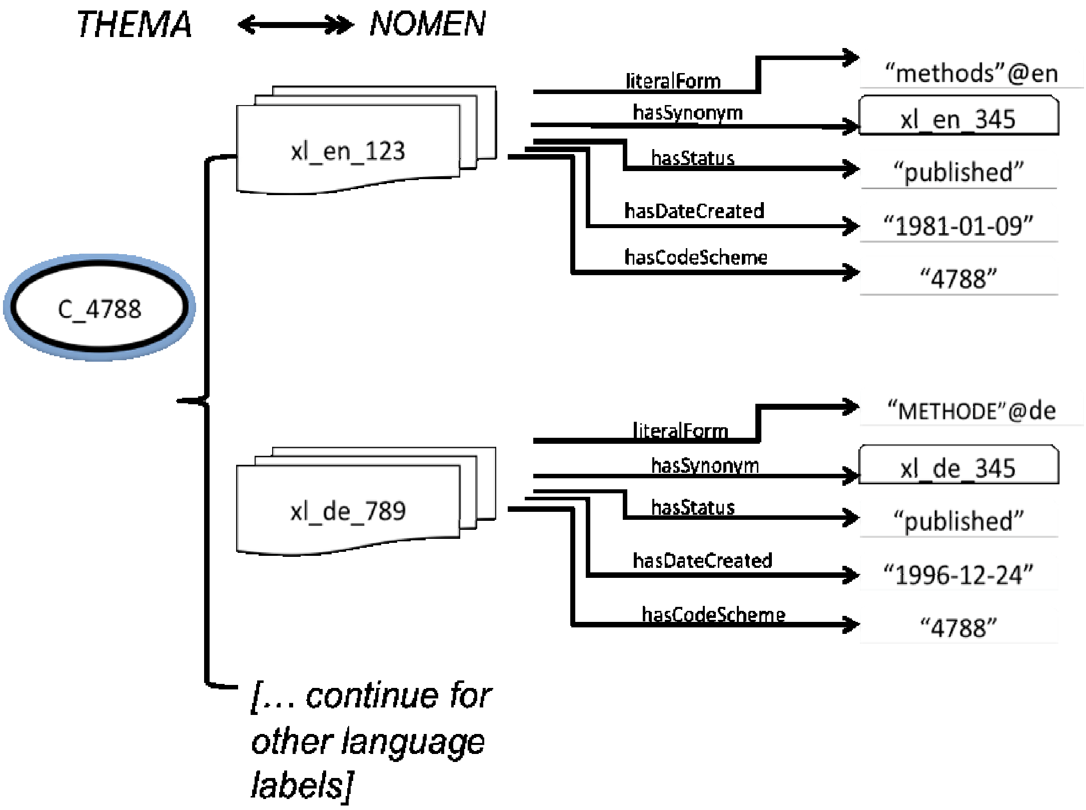


Figure 7. Simplified illustration of a multilingual thesaurus entry data model, where Label is an entity and has attributes.

The data model supporting this structure can be summarized as the following:

1. Each concept has multiple preferred labels in different languages. (Each language has one preferred label.)
2. Each concept also has one or more non-preferred labels of each involving language. They are considered to be synonyms of the preferred label.
3. Multilingual preferred labels have different literal forms, synonyms, status, creation date, and other administrative information.
4. Label has attributes, e.g., 'hasLiteralForm', 'hasStatus', 'hasDateCreated'.
5. The label instance may have a relationship with another label instance, which is demonstrated by the "'has synonym' relationship between label "xl_en_123" and label "xl_en_345". These would correspond to FRSAD Nomen equivalence relationship.

With SKOS, all original functions for Concept class still apply (e.g., for presenting the established semantic relationships between concepts, the attributes of concepts, and for organizing concepts in a concept schema, aggregating, and mapping concepts from different schemas) and these are also applied to the Label class.

While the most common relationships and attributes are specified in FRSAD, it is made clear that the list in not prescriptive: additional implementation- or domain-specific relationships and attributes can be added when needed. The two data models presented above demonstrate how the FRSAD model can be implemented as well as the power that such a model wields to meet the needs of both the conventional LIS environment and the emerging Linked Data environment.

References

Dextre Clarke, S. G. (2008). ISO 2788 + ISO 5964 + Much Energy = ISO 25964. *Bulletin of the American Society for Information Science and Technology*. 35(1), 31–33.

Food and Agriculture Organization of the United Nations. *AGROVOC Thesaurus* (Linked Data version 2011-). Retrieved from: http://aims.fao.org/website/Linked-Open-Data/sub

IFLA Working Group on Functional Requirements for Subject Authority Records (FRSAR) (2010). *Functional Requirements for Subject Authority Data, A Conceptual Model* . Retrieved from: http://www.ifla.org/en/node/1297

Isaac, A. and Summers, E. (Eds.) (2009). *SKOS Simple Knowledge Organization System Primer*. Retrieved from: http://www.w3.org/TR/skos-primer/

Library of Congress. *Library of Congress Subject Headings* (Linked Data version 2009-). Retrieved from: http://id.loc.gov/authorities/

Miles, A. & Bechhofer, S. (Eds.) (2005*). SKOS Core Vocabulary Specification. W3C Working Draft 10 May 2005*. Retrieved from http://www.w3.org/TR/2005/WD-swbp-skos-core-spec-20050510/

Miles, A. & Bechhofer, S. (Eds.) (2009*). SKOS Simple Knowledge Organization System: reference. W3C Recommendation 18 August 2009*. Retrieved from http://www.w3.org/TR/2009/REC-skos-reference-20090818/

Miles, A. & Bechhofer, S. (Eds.) (2009*). SKOS Simple Knowledge Organization System: reference. W3C Recommendation 18 August 2009*. Appendix B. SKOS eXtension for Labels (SKOS-XL). Retrieved from http://www.w3.org/TR/2009/REC-skos-reference-20090818//#xl