XKOS - An RDF Vocabulary for Describing Statistical Classifications

by Franck Cotton, Daniel W. Gillman, Yves Jaques[1]

Introduction

This paper contains a brief description of the eXtended Knowledge Organization System (XKOS) and a rationale for why it was developed. In particular, there is a focus on describing statistical classifications with XKOS. For statistical data, statistical classifications are essential for categorizing complex domains, such as industries or occupations; presenting dimensions on which to aggregate data, such as in tables or time series; providing the means to stratify populations; and supplying survey respondents with standard response choices.

XKOS is an extension of the Simple Knowledge Organization System (SKOS)[2] applicable to the needs of statistical offices and social science data users. As we show in this paper, some limitations in SKOS leave it inadequate to the task of describing statistical classifications. XKOS is designed to fill these gaps.

SKOS was published in 2009 as a World Wide Web Consortium (W3C)[3] recommendation, and in the same year was extended in another vocabulary named SKOS-XL. This was to better meet the needs of multi-lingual thesauri. The purpose of SKOS is to provide a representation for knowledge organization systems, of which statistical classifications and thesauri are examples, in a machine-understandable way within the framework of the Semantic Web[4]. Therefore, SKOS-encoded statistical classifications are appropriate for use within the Linked Open Data (LOD)[5] community.

LOD is a set of recommendations for building the Semantic Web, described by Tim Berners-Lee in 2006,[6] and has been taken up by a wide variety of communities including biodiversity, environment, statistics, GIS, libraries, archives, and museums. Its promise to provide crosswalks across domains and types of data is especially attractive to the growing "open access" and "open data" movements that in the social science data community are beginning to force change to the business-as-usual practice of considering each dataset part of its own closed world.

The purpose of SKOS is to provide a representation for knowledge organization systems

Implementing the LOD recommendations provides new abilities to find, understand, and combine data on similar or otherwise related domains by organizing and linking data and metadata. LOD adds value to disparate, difficult to link datasets by employing frameworks such as the Resource Description Framework (RDF).[7] RDF, described further in the

Resource Description Framework section, is a W3C standard used for organizing and linking data. Links are used to navigate and find related data and metadata; therefore the technique, among other features, provides an easy to leverage mechanism for building mash-ups (data from multiple sources).

Implications of using LOD for data harmonization were initially explored in a paper by Gillman (2010[8]), which includes references to work to mash-up crime, traffic, workplace safety, and natural disaster risk data to create a livability index for US cities. Even though the cited work did not employ LOD *per se*, the ideas are very similar to LOD recommendations, and the reader is encouraged to understand the example. Moreover, the example shows that to do LOD right in the statistical framework is not at all straightforward. However, as a growing collection of new tools and many applications have been built with LOD, there is an expanding community of interest in employing the technology, and many benefits are promised.[9] The statistical data community needs to be paying attention to these developments.

Along with XKOS, other RDF developments that affect the statistical data community have taken place. The Data Cube vocabulary built through cooperation between LOD experts and SDMX technical experts has produced a rendition of SDMX for LOD[10] which is already in wide use by major initiatives such as data.gov.uk.[11] Similar work is planned for DDI, and the workshops held at Schloβ Dagstuhl[12] in Germany on *Semantic Statistics for Social, Behavioural, and Economic Sciences: Leveraging the DDI Model for the Linked Data Web* in September 2011[13] and October 2012[14] were devoted to the topic. In particular, this is where XKOS was first developed.

The original SKOS is used widely in LOD applications, as seen in the SKOS Implementation Report.[15] As a result, a group was formed at the Dagstuhl Workshops (in 2011 and 2012) to look at the suitability of using SKOS in the statistical data community for LOD work. As will be described in the SKOS / What is Missing section, SKOS was found to have shortcomings, so the group looked to address the issues of how to extend SKOS to meet the needs of the statistical data community. Several extensions were deemed important enough for inclusion under a new initiative, XKOS, with the intention of submitting this as a W3C Editor's Draft. Fortunately, the entire design and culture of RDF is based on a spirit of re-use and extension, so extending SKOS is technically easy. The results of the workshops and subsequent output are reported here.

In this paper, we provide introductory remarks to set the stage for discussion, provide a short primer on RDF, describe SKOS in general and the limitations to statistical classifications embedded in the design in some detail, and lay out the extensions to SKOS that form the XKOS specification. In particular, we show how the semantics of classification systems in our own offices are represented more faithfully by extending SKOS with XKOS.

Resource Description Framework

This section gives a brief primer on RDF, a W3C standard that facilitates the exchange of structured data on the Internet. Based on a simple subject-predicate-object model commonly referred to as "triples," it allows for a generic, standardized structuring of resources that can be used to model and disseminate everything from taxonomies to statistical observations to metadata records. The model used by RDF is also commonly referred to as a "graph model" consisting of "nodes" (which are vertices) and "edges" or "arcs." See the Figure 1 below for an example.

The RDF model, which by itself contains only the barest set of classes (subjects and objects) and properties (predicates), is extended using RDF Schema,[16] another fairly limited set of classes and properties that together with RDF form the foundation of the framework which can then be endlessly extended and specialized as needed. Each extension is known as a *vocabulary*, which is bounded by a *namespace*. Namespaces allow implementers to specify the *set* of classes and properties that belong to a vocabulary and give a strong assurance of uniqueness even in the open waters of the World Wide Web (WWW). This is a concept that will be familiar to those who know XML schemas.

The other very important aspect of RDF is that as with its namespaces, all of its classes and properties are also uniquely identified using the underpinning naming mechanism of the Internet, the URI[17] (Uniform Resource Identifier). In the same way that all web pages are uniquely identified by a URI (web pages actually use the URL,[18] a subset of the URI specification), all RDF classes and properties are uniquely identified by a URI. In practice this enables a powerful, standardized method for uniquely identifying information of all kinds with great certainty that the information will remain unique not only within the closed context of an internal database, but also across the WWW.

As mentioned before, each vocabulary uses a namespace to scope its set of classes and properties. This namespace is known by a URI, and by common convention the unique identifiers for the classes and properties are appended to this common namespace URI with an intervening hash or forward slash. For example, the commonly used Friend of a Friend (FOAF)[19] vocabulary, designed to link instances of people and information, uses the common namespace *http://xmlns.com/foaf/0.1/*. All of the FOAF classes and properties are then appended to this namespace, e.g., the FOAF class Person is uniquely identified by its URI as *http://xmlns.com/foaf/0.1/Person*.

Just as in an XML schema, one can define a namespace prefix to act as a shortcut for the entire namespace. Thus in a group of FOAF statements (written in XML syntax) one will commonly find a statement such as *xmlns:foaf=http://xmlns.com/foaf/0.1/*. This simply means that once this foaf shortcut has been defined, one can now refer to the URI that uniquely identifies the class FOAF Person more compactly as *foaf:Person*.

One of the other important aspects of RDF is that it does not rely on a particular syntax for its expression. Thus, there are a handful of interchangeable syntaxes that can and are used depending on a variety of requirements that one may have such as brevity or readability. This paper uses the popular Turtle (Terse RDF Triple Language[20]) syntax, prized for its readability.

Returning to the FOAF example, here is how one might make the simple triple statement that one of the authors of this paper is a thing known as a person (with a web page to provide an identifier for the actual person):

< http://aims.fao.org/community/profiles/yjaques>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://xmlns.com/foaf/0.1/Person>.

So to recap, we have a subject "Yves Jaques", a "type" predicate (defined in RDFS), and an object *foaf:Person*.
To put it in another way, "Yves Jaques" is an instance of the class "Person". In RDF "type" gets used so often that Turtle lets you simply use "a" for convenience:

< http://aims.fao.org/community/profiles/Yves-Jaques>
a
<http://xmlns.com/foaf/0.1/Person>.

Let's say we want to make our statement a little shorter. We can define namespace prefixes one time and then use the shortcut for all the other triples in our graph:

@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix aims: <http://aims.fao.org/community/profiles/> .

So with those shortcuts defined, we can now write the same statement as (putting the triple on a single line this time):

aims:Yves-Jaques rdf:type foaf:Person .

Or using the Turtle shortcut for rdf:type:

aims:Yves-Jaques a foaf:Person .

Let's say we want to put a few triples together so we can say a little bit more:

aims:Yves-Jaques
a foaf:Person ;
foaf:name "Yves Jaques" .

So here we are seeing the short-hand Turtle notation for two sets of triples. In words, these triples are

"The Yves-Jaques AIMS profile web page is a person."
"The person is named Yves Jaques."

This illustrates another feature of RDF. The triples may be linked together to tell a story. The object in the first triple is then used as the subject in the next (possibly many) triple(s).

To think about what RDF looks like graphically, here is a nice diagram courtesy of Marek Obitko.[21] The round-cornered boxes are classes or instances of classes (subjects/objects), the arrows are properties (predicates), and the square boxes are *literals*. Literals are typically used to represent simple numeric values, dates, or labels. Literals can also have a datatype, a powerful mechanism to enforce restrictions on permissible values:

And here is the corresponding Turtle (note the use of the empty namespace shortcut):

@prefix : <http://www.example.org/~joe/contact.rdf#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

:joesmith a foaf:Person ;
foaf:givenname "Joe" ;
foaf:family_name "Smith" ;
foaf:homepage <http://www.example.org/~joe/> ;
foaf:mbox <mailto:joe.smith@example.org> .

To briefly recap, RDF is a framework that is designed to organize structured data about resources and their relationships over the Internet in a standard way. It is designed from the ground-up to be endlessly extensible and able to maintain the uniqueness of the things it represents even in the radically decentralized WWW.

SKOS

What Is Missing

SKOS provides a means for representing knowledge organization systems using RDF, and this makes the use of SKOS immediately applicable to LOD and the Semantic Web. So, SKOS is important for

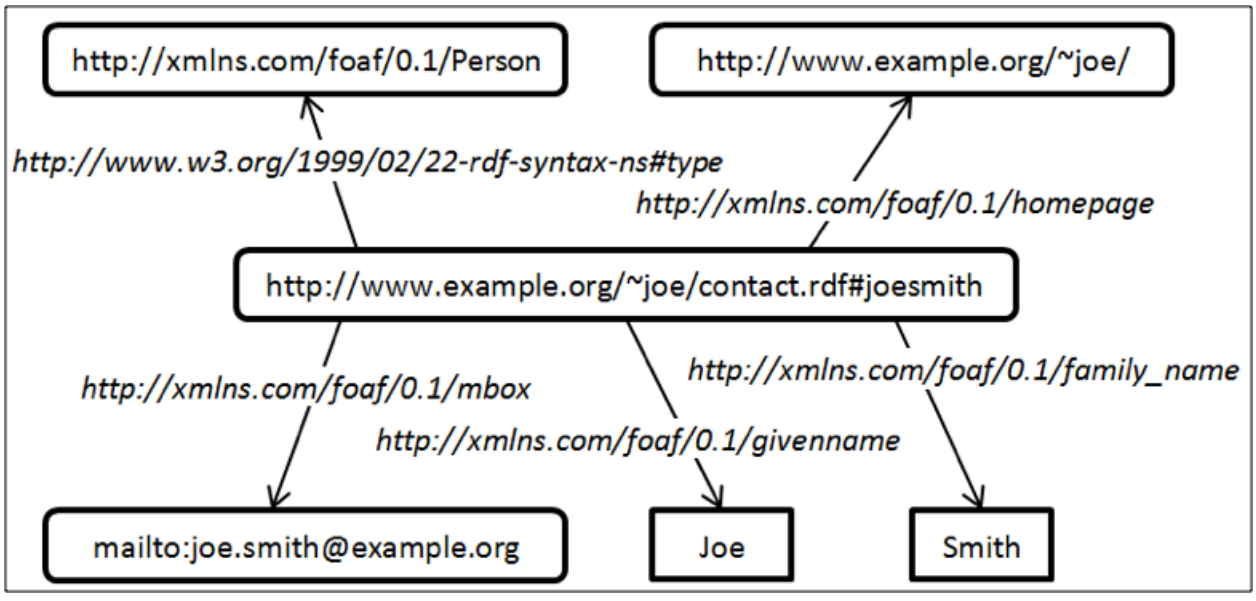


Figure 1: RDF Graph

organizations that wish to use LOD and employ classifications and code sets.

It is beyond the scope of this paper to provide a detailed description of SKOS. We direct the interested reader to the SKOS website (see End Note 2). However, SKOS contains the following basic ideas, whose definitions we paraphrase here:

- Concept Scheme – any knowledge organization system (including statistical classifications and code sets)
- Concept – any abstract idea or unit of thought
- Definition – formal statement conveying the meaning of a concept
- Label – lexical representation for a concept, may be preferred or alternate; provides means to communicate the concept
- Notation – a symbolic notation for the concept (such as a code) that is typically data-typed
- Semantic Relation – broad category for relations between concepts, such as broader than, narrower than, and related to (these relations can include relations to concepts found in other concept schemes)

The basic ideas listed above are the minimum required to describe a classification scheme. We can account for the scheme itself (*concept scheme*), all its underlying concepts (or categories as they are often called in statistics) with *concept*, what each concept means (*definition*), the labels and codes associated with a category (*label / notation*), and relationships between a concept and its parent and between and concept and all of its children (*semanticRelation*). So, is anything missing that is needed for statistics?

SKOS is based on the now withdrawn standard ISO 2788 - *Guidelines for the establishment and development of monolingual thesauri.* This standard describes three basic kinds of relations between concepts: *generic, partitive,* and *instantiation*. The *generic* relation refers to a generic / specific situation, such as between family and genus/species in the biological classification of living things. For instance, all *Homo sapiens* are mammals. The *partitive* relation refers to a part / whole situation, such as between an automobile and a steering wheel. *Instantiation* is the relation between a kind and an instance, such as each of the authors of this paper are instances of the class of people. Both the *generic* and *partitive* relations are used in statistical classifications, but *instantiation* is not.

Interestingly, the *generic* and *partitive* relations are not provided in SKOS, only the more generic *broader than* and *narrower than*, which are often referred to in more technical settings as *super-ordinate* and *sub-ordinate*, respectively. Both the *generic* and *partitive* relations are specializations of *broader than / narrower than*. In the SKOS Primer,[22] this simplification is acknowledged by the following:

> "Not covered in basic SKOS is the distinction between types of hierarchical relations: for example, instance-class and part-whole relationships. The interested reader is referred to Section 4.7, which describes how to create specializations of semantic relations to deal with this issue."

These more specialized relations were included in the past in SKOS, but they are now deprecated. XKOS, in part, is the effort to put them back.

SKOS also specifies the possibility of an *association* relation between concepts, but this is not made any more detailed. It is possible to specialize *associations* somewhat, and that is done in XKOS through *sequential, temporal*, and *causal* relations, none of which are in SKOS. The *sequential* relation refers to ideas where one is the antecedent of the other, either temporally or spatially. An example is the relationship between production and consumption. The specialized *temporal* relation is based on time. An example is the relationship between spring and summer. Finally, the *causal* relation relates cause and effect, such as the detonation of a hydrogen bomb and nuclear fall-out. Upon inspection of some classification schemes in the statistical offices of the authors, some of these relations are needed.

There is also a structural deficiency in SKOS; there is no satisfactory way to represent the idea of levels in concept schemes. Levels in statistical classifications are used to identify aggregation levels in reported statistics, which provide producers a consistent way to report their data or provide a way to reduce the threat of disclosures. Therefore, XKOS also needs to account for levels in concept schemes.

Examples

Below are some examples that illustrate the need for the extensions we have identified above:

1. The US Standard Occupational Classification System (SOC – 2012).

Take, for example

27-2000 – Entertainers and Performers, Sports and Related Workers
27-2040 – Musicians, Singers, and Related Workers
27-2042 – Musicians and Singers

The appropriate relation between 27-2000 and 27-2040 is generic, i.e. Musicians, Singers and Related Workers is a specialization of Entertainers and Performers, Sports and Related Workers. The same relation is found between 27-2040 and 27-2042, i.e., Musicians and Singers is a specialization of Musicians, Singers and Related Workers. So, the *generic* relation is needed to specify the semantics of the US SOC.

2. The US Occupational Injury and Illness Classification[24] (OIICS – 2012).

Occupational injury and illness is a four-facet classification: nature, body part, source, and event. In the body part facet, for example

3 – Trunk
31 – Chest
313 – Heart
315 – Lungs
32 – Back, including spine, spinal cord
321 – Thoracic
322 – Lumbar

Going from broad to lower detail in this snippet of the body part classification illustrates the *partitive* relation. The chest and back are parts of the trunk. The heart and lungs are part of the chest. Finally, the thoracic and lumbar regions are part of the back and spine. Note that it would not be proper to use the *generic* relation here. Therefore, the partitive relation is needed to specify the semantics of the US OIICS.

3. The US American Time Use Survey — Activity Coding Lexicons,[25] last updated in 2011. The classification is a hierarchy, but some activity categories depend on what has occurred before. For instance,
04 – Caring For & Helping non-Household Members
0402 – Caring For & Helping non-Household Children
040204 – Arts & Crafts with non-Household Children
040212 – Dropping Off/Picking Up non-Household Children

Dropping off non-household children is a sequential activity related to having supervised arts-and-crafts activities (or some other activity in the 04 group) previously. So, there are associations between some pairs of activities within this classification. In this case, the sequential or possibly the temporal relation is needed to convey the additional semantics that some activities depend on the triggering of other prior activities.

XKOS

We move now to a description of the XKOS vocabulary. As already mentioned, just as SKOS-XL extends SKOS for the needs of multi-lingual thesauri, XKOS extends SKOS for the needs of statistical classifications. It does so in two main directions. First, it defines a number of terms that allow the representation of statistical classifications with their structure and textual properties, as well as the relations between classifications. Second, it refines SKOS semantic properties to allow the use of more specific relations between concepts. Those specific relations can be used for the representation of classifications or for any other case where SKOS is employed.

Classifications

For the representation of statistical classifications, XKOS borrows from the Neuchâtel Model,[26] which is a *de facto* standard created by a group of statistical institutes and maintained in the United Nations Economic Commission for Europe's Common Metadata Framework.[27] XKOS is not a complete translation of the model, though. In particular, the notion of a classification index is not supported. There are other areas where minor differences exist between the XKOS and Neuchâtel Model approaches: these will be described below.

To begin with the classification itself, we distinguish within XKOS the notion of classification and that of classification scheme. A classification is a set of classification schemes that share a well-known name, for example, the European Statistical Classification of Economic Activities (NACE) or the International Standard Industrial Classification (ISIC). Typically, a classification scheme will be a *major version* of a given classification. For example, NACE is a *classification*, and each version of NACE (the original 1970 version, the 1990 NACE Rev. 1, the 2003 NACE Rev. 1.1, and the 2008 NACE Rev. 2) are *classification schemes* belonging to this classification.

The Neuchâtel Model also defines the Classification Variant, which is an adaptation of a classification version to a certain context or usage. In a variant, items can be split, aggregated, added, or suppressed relative to the standard structure of the base version. A variant can also be represented as an XKOS Classification Scheme, albeit of a particular type.

XKOS does not create its own object classes to represent classifications, classification schemes, and classification items, but directly uses classes already defined in SKOS. Classification items will be represented as instances of *skos:Concept*, with normal SKOS properties for codes, labels, etc. A classification scheme will simply be a *skos:ConceptScheme*, which is defined as an aggregation of concepts and semantic relationships between those concepts. A classification itself will also be a *skos:Concept*, which can in turn be included in concept schemes representing classification families (e.g., "Occupational classifications", "Activities classifications", etc.).

However, XKOS defines a set of properties that can be used to link classifications and classification schemes. For example, *xkos:belongsTo* allows one to attach a classification scheme to its classification, and *xkos:follows* or its sub-property *xkos:supersedes* can link classification schemes representing successive versions of a classification. XKOS also provides a set of properties that indicate how a classification covers its field (e.g., exhaustively, without overlap, both). The field itself would be a SKOS concept that can be taken from a well-known thesaurus such as Eurovoc[28] or the Library of Congress Subject Headings.[29]

Of course, existing standard RDF properties are available to capture versioning information, textual documentation, etc. Examples of these are the Dublin Core[30] *dcterms:valid* property, or the RADion[31] *radion:version* property. Also, *skos:note* can be used to record documentation or other descriptive resources relative to classifications and schemes. In keeping with the RDF spirit of re-use, the existing classes and properties of broadly supported vocabularies are used wherever possible.

The main purpose of a classification is to classify the entities that belong to or operate in the field that it covers. In linked data terms, classification results in the creation of an RDF triple where the subject is the resource representing the entity and the object is the concept representing the classification item. XKOS defines a generic property, *xkos:classifiedUnder,* that can be used in such statements, but classification criteria are often quite complex: for example, the same enterprise could be classified in different items of a classification of activities, depending on the rules that are used to measure its main economic activity. Thus, it is expected that *xkos:classifiedUnder* will be specialized for use in specific contexts.

Another important notion in the classifications terminology is the notion of level. Many statistical classifications, especially those that are international standards, are organized in embedded levels. For example, the ISIC Rev. 4 has four levels: the top is composed of 21 sections that cover broad economic sectors, and there are three more levels that go into greater and greater detail: divisions, groups, and classes.

In SKOS terms, classification levels are just *collections* or at most *ordered collections* of concepts, but their hierarchical organization within a classification scheme gives them extra characteristics not covered by SKOS. Thus, XKOS defines a dedicated subclass of *skos:Collection* to represent them, which is the *xkos:ClassificationLevel.* The levels or instances of *xkos:ClassificationLevel*, are structured as an RDF List, starting with the most aggregated, and the list is attached to the classification scheme by the *xkos:levels* property. An *xkos:depth* property can be used to express the distance of a given level from the (abstract) root node of the level hierarchy, and an *xkos:organizedBy* property

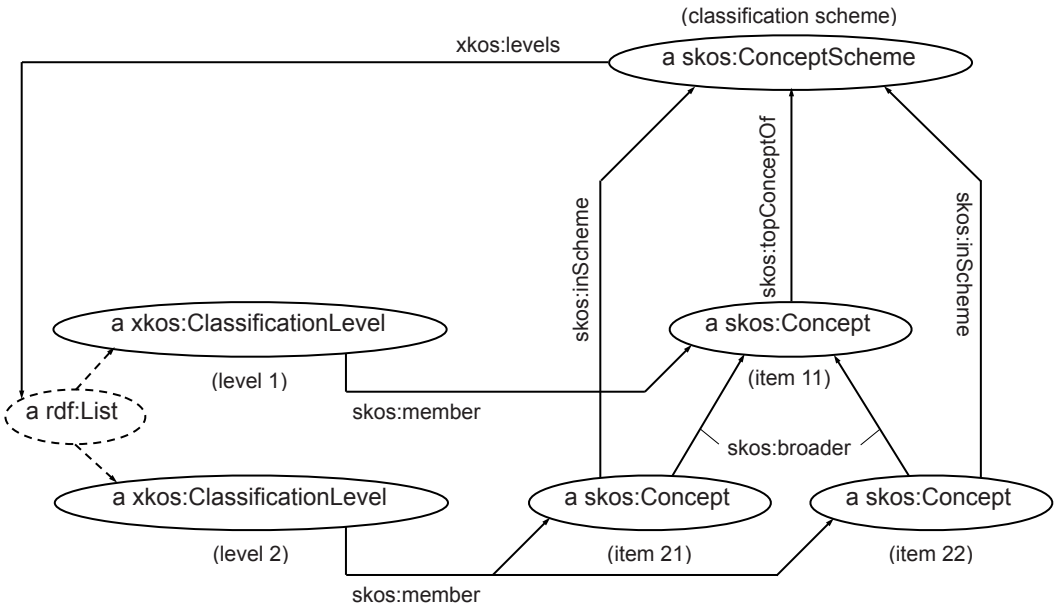


Figure 2: Structure of a Classification Scheme

can be used to record the generic name of the items of a given level (e.g., "section", "division", etc.).

The structure of a classification scheme can be described using the usual SKOS properties. More precisely:

- *skos:inScheme* (or the more specific sub-property *skos:topConceptOf* if the items belong to the most aggregated level) links the classification items to the classification scheme
- *skos:member* connects the classification level to the items that it contains
- *skos:broader* and *skos:narrower* represent the hierarchical relations between the classification items

In this last case, the more precise sub-properties defined by XKOS to express partitive or generic relations between concepts (see below) may be used instead of *skos:narrower* or *skos:broader*.

Figure 2 illustrates a simple abstract case of the usage of SKOS properties to represent the structure of a classification scheme.

Textual properties

Good classifications usually come with a fair amount of textual material, generally organized as notes attached to the classification items or to the scheme itself. These notes typically explain the content of a given classification item by describing what should be classified under this item and what should go elsewhere.

For example, here is an excerpt from the official publication of NACE:[33]

46.34 Wholesale of beverages

This class includes:

- wholesale of alcoholic beverages
- wholesale of non-alcoholic beverages

This class also includes:

- buying of wine in bulk and bottling without transformation

This class excludes:

- blending of wine or distilled spirits, see 11.01, 11.02

We see that the explanatory notes have a defined structure: they first describe what is included in the item, then what is excluded. For the inclusions, a distinction is made between what is evidently included (sometimes called "central content" or "core content"), and what is "also" included, by convention or experts' decisions, even if it does not result obviously from the item's label. For the exclusions, the note often refers explicitly to the item(s) where the content should in fact be classified.

It is perfectly satisfactory to represent explanatory notes with SKOS generic notes (*skos:note*) or scope notes (*skos:scopeNote*), but it can be useful to be able to easily distinguish between the different types of note. For this purpose, XKOS introduces four sub-properties of *skos:scopeNote*, which are represented in the Figure 3 below.

In the case of the NACE class 46.34 cited before, we would have three RDF triples to represent the explanatory notes with predicates, respectively, *xkos:coreContentNote*, *xkos:additionalContentNote* and *xkos:exclusionNote*. SKOS does not specify which type the objects of these triples should be, nor does XKOS. As a side note, Eurovoc uses an interesting mechanism that allows the representation of the notes as XHTML fragments, thereby opening the possibility of rendering the references to other items as HTML links.

Correspondences between classifications

Different classification schemes can cover the same classification, the same field, or even fields that are different but semantically related. This induces semantic relations between the classification items that belong to these schemes. A simple example of this is given by two successive major versions of a classification: some items may remain unchanged in the new version, but others will disappear, merge, be created, etc. More complicated n to m correspondences between items of the two versions are frequent.

A much more complex example of relations between classifications or classification schemes is given by the international system of economic classifications maintained by

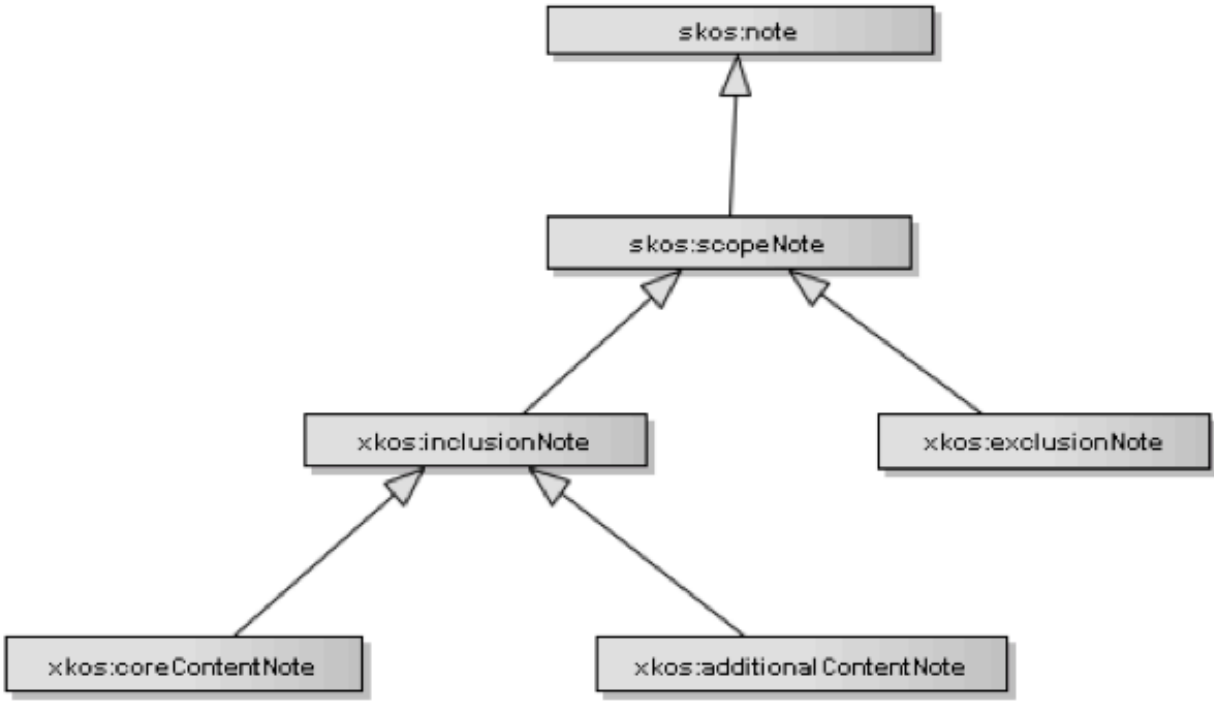


Figure 3: XKOS Note Properties

the United Nations Statistical Division. The European view of this system is well described in the online publication of the NACE Rev. 2 (*op. cit.*, chapter 1.1). The economic classifications forming this system are linked either by a common structure which gets more detailed as one goes from the international to the European to the national levels, or by semantic correspondences between the economic fields covered: activities, products, and goods (e.g., activities create products). Here again, the high-level links established between classifications result in more fine-grained correspondences between items: a given activity will create one or more specific products.

Thus, there are different types of correspondences between classifications, schemes, or items:

- Between classifications on the same field, for example, North American and European activities classifications
- Between different linked fields, for example, classifications of activities and products
- Historical correspondences, for example, SIC to NAICS
- Versioning of items over time within a given classification scheme

Since classification items are represented as SKOS Concepts, we could use the usual SKOS associative properties to represent

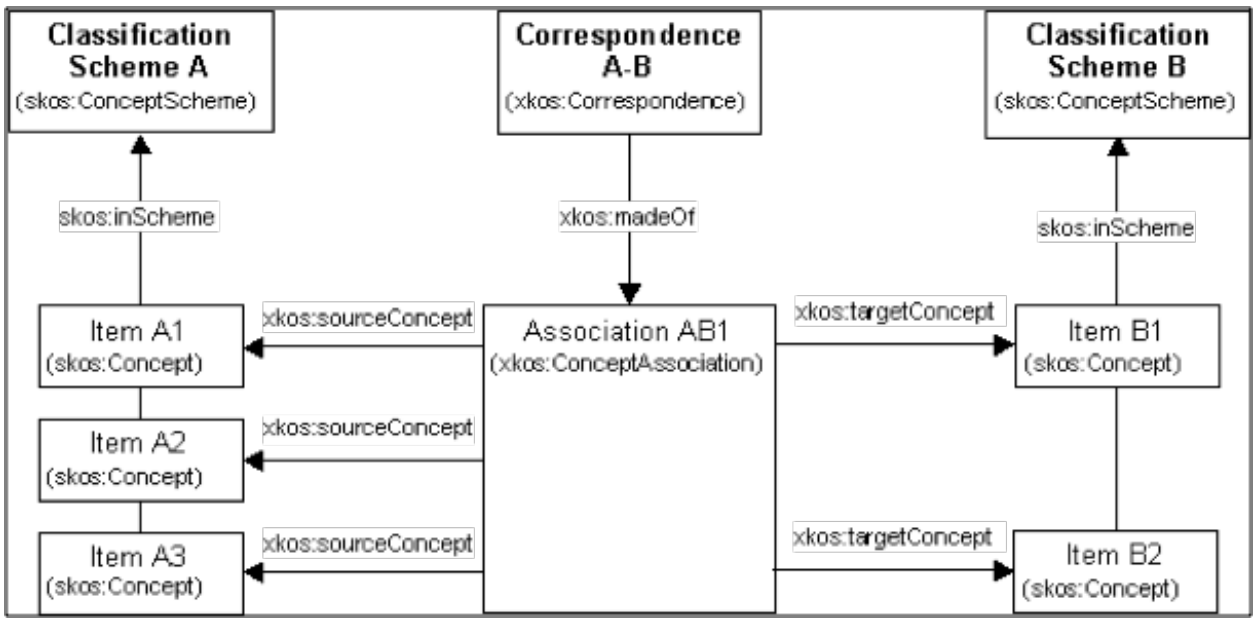


Figure 4: Concept Association Example

correspondences between them. However, this simple approach has some limitations:

- As mentioned above, relations between items in correspondences are often n to m, whereas SKOS properties relate one unique concept to another unique concept. It is always possible to decompose an n to m relation into several *1 to 1* relations, but it is better to have a global vision of a given correspondence. We also want to be able to represent *0* to *n* relations, for example, when an item is created or disappears in a new version of a classification.
- More globally, we want to be able to group all the fine-grained item associations that compose a given high-level relation between two classification schemes, such as the ones that exist in the international system of economic classifications. Such a collection of item associations is called a correspondence table, conversion table, or concordance.
- Lastly, it is often useful to be able to attach additional information (for example, notes) to item associations, for example, to describe what proportion of the different items are linked in the association.

For these reasons, XKOS defines the *xkos:ConceptAssociation* class that can be used to represent correspondences between classification items where the SKOS properties are not sufficient. Each *xkos:ConceptAssociation* may have input or source *skos:Concept*(s) and output or target *skos:Concept*(s). The complete collection of such associations for all the concepts in two SKOS Concept Schemes forms a correspondence and is expressed as an instance of the *xkos:Correspondence* class. The *xkos:madeOf* property is used to link the *xkos:Correspondence* to its *xkos:ConceptAssociation* components. To those familiar with entity-relationship diagrams, what XKOS does is to take the *skos:related* relationship (property) and "decompose" it into its own entity (class) to solve the n to m relationship problem as well as to be able to add additional properties to the relationship.

Figure 4 illustrates a simple example of a concept association: three classification items are re-combined into two.

The *xkos:ConceptAssociation* is similar to the Correspondence Item in the Neuchâtel model, but it can describe in a single instance the relationship of any number of source concepts to any number of target concepts rather than expressing the association through a set of pair-wise relations. The XKOS concept association can also represent the *Item Change* class of the Neuchâtel model. However, in this version, XKOS does not define any properties or sub-classes for *xkos:Correspondence* and *xkos:ConceptAssociation* for modeling the different types of correspondences that we described above, nor can XKOS describe the typology of item changes detailed in the Neuchâtel model (Annex 3). These may be added in a future version.

Semantic properties

Semantic properties constitute the second direction in which XKOS extends SKOS. Concept schemes are not just lists of concepts: as the SKOS Primer puts it (section 2.3), "The meaning of a concept is defined not just by the natural-language words in its labels but also by links to other concepts in the vocabulary."

SKOS intentionally defines few properties, but introduces the fundamental distinction between hierarchical and associative relations. In both these categories, XKOS creates more precise properties which are described below. The reader can refer to the figure provided in Annex 1 to find a panoptic view of SKOS and XKOS properties.

Hierarchical properties

SKOS defines several hierarchical properties, but the most used are *skos:broader* and *skos:narrower*, which are each other's inverse. These are the two properties that are refined in XKOS. A concept is broader than another one if it encompasses a wider portion of the field covered by the concept scheme, and thus includes the scope of the narrower concept. Note that the *skos:broader* property has the narrower concept for the subject and the broader one for the object, for example, "Car""broader""Vehicle"; and the *skos:narrower* property has the broader concept for the subject and the narrower one for the object, for example, "Green""narrower""Olive".

As we made clear in the previous sections, it is important, at least for statistical purposes, to represent generic and partitive relations between concepts. XKOS therefore defines two couples of inverse properties: *xkos:specializes* and *xkos:generalizes* on the one hand, *xkos:isPartOf* and *xkos:hasPart* on the other. All are sub-properties of *skos:broader* and *skos:narrower*, but the terminology is a bit tricky here: *xkos:specializes* goes from the more specific concept to the more generic one, and thus is a sub-property of *skos:broader*. Similarly, *xkos:hasPart* is a sub-property of *skos:narrower*. For example, head *isPartOf* body and chest *hasPart* heart.

Associative properties

In terms of associative properties, SKOS defines the very general *skos:related,* and a set of mapping properties (*skos:closeMatch, skos:exactMatch*, etc.) intended for establishing links between concepts of different schemes. XKOS proposes a hierarchy of *skos:related* sub-properties that convey more precise semantics. This hierarchy is organized in three branches.

The *xkos:disjoint* property forms a branch of its own. In some circumstances, it is useful to explicitly state that two given concepts do not overlap (for example, *private company* and *non-profit organization* in the *Class-of-Work* classification of the *US Current Population Survey*), especially when it has not been specified that the scheme covered its field without overlap (see A.1 in figure 4 above).

The second line of XKOS associative properties is dedicated to causal relationships. This class of link between concepts is frequently encountered (physics, biology, history, law, etc.). The generic *xkos:causal* is further subdivided into *xkos:causes* and *xkos:causedBy*, so that the direction of the causality can be expressed.

The last branch of properties is the most populated and deals with sequential relationships; it is represented on Figure 5 below. The top node of this branch is *xkos:sequential*, a refinement of *skos:related* that just indicates that two concepts in a scheme are in a sequential relationship, for example, notes in a musical scale. Below are *xkos:succeeds* and *xkos:precedes* that can be used when the sequence has a known order between the concepts. A third sub-property of *xkos:sequential* is *xkos:temporal*, which can be used when the sequence is of a temporal nature (i.e., events in time.). *xkos:temporal* itself is the parent of *xkos:before* and *xkos:after*.

It was found useful to add two more precise sub-properties of *xkos:precedes* and *xkos:succeeds*, namely *xkos:previous* and *xkos:next*. Previous and next imply that there is no intermediary concept

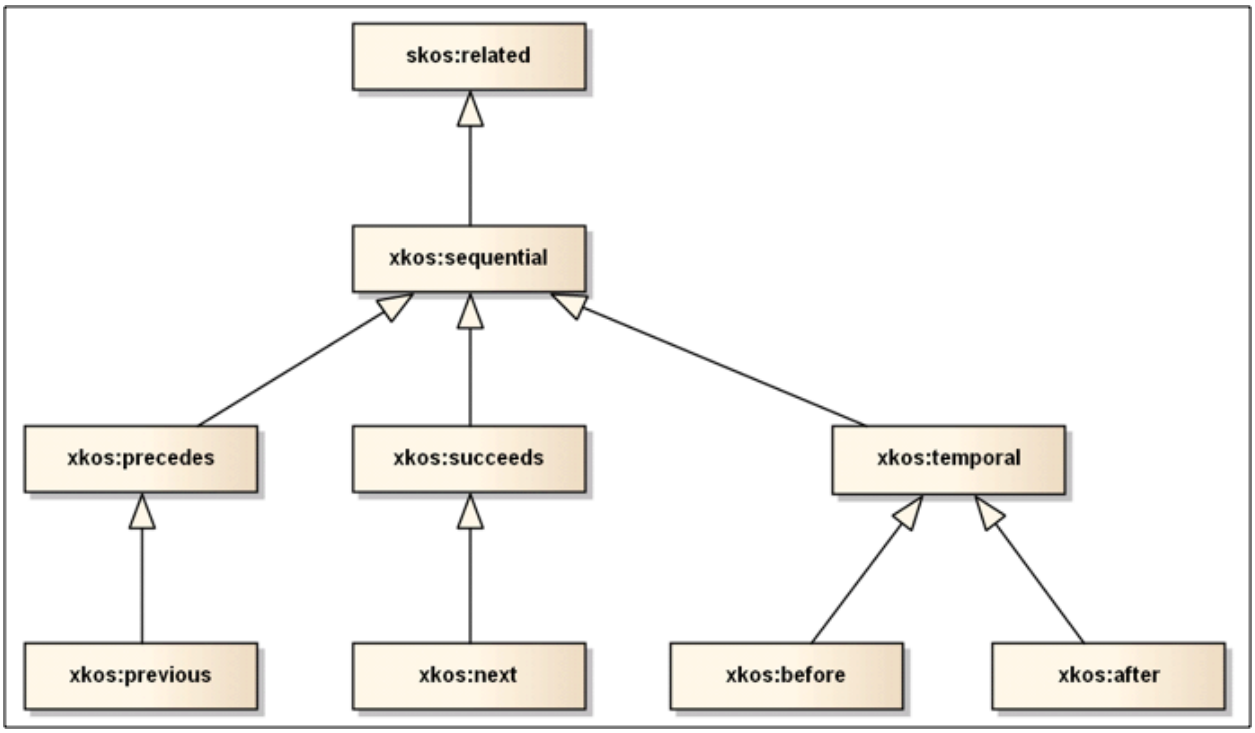


Figure 5: XKOS Sequential Properties

between two sequentially linked concepts. These two properties are of course not transitive, although their parents are.

Conclusion

In this paper, we laid out the general rationale and purpose for why XKOS was developed. We explained the basic extensions to SKOS that were identified as needed to describe statistical classifications in the LOD domain, and we gave examples from the statistical community to justify our choices. Some unresolved issues were also discussed. Finally, we gave a rationale for the importance of SKOS and XKOS, appealing to the burgeoning LOD community of practice, the use of RDF, and the growth of the Semantic Web in general.

It is interesting that some of the extensions (*generic* and *partitive* relations) were originally included in SKOS. Given the amount of discussion in the LOD and Semantic Web communities about semantics and precision, it is even more remarkable that these specific relations were left out. On the other hand, there was a clear desire by the SKOS designers to make building Semantic Web applications as simple as possible. Since SKOS is the Simple Knowledge Organization System, this design choice begins to make sense.

Yet, we have also seen that the worlds of thesauri and classifications are often too complex to model in SKOS. Thus, the vocabulary was quickly extended with SKOS-XL to handle the need to treat labels, not as literals, but as actual class instances (a process sometimes referred to as *reification*) that could participate in relationships with other instances and have properties of their own. While SKOS-XL extends SKOS for the particular needs of the multi-lingual thesaurus community, XKOS adds the extensions that are desirable to meet the requirements of the statistical community.

SKOS is a very popular specification, and we hope the XKOS extensions will simply serve to increase its adoption. The proof of whether XKOS is useful will be found when statistical offices implement it. This work is already underway. However, XKOS is still a work in progress, and unresolved issues remain. We hope the users of XKOS will offer help with these issues, provide comments to the authors on the effectiveness of XKOS, and give guidance as to what other areas should be extended as we prepare to submit the standard as a W3C Editor's Draft.

Acknowledgements

The authors wish to thank the organizers of the Dagstuhl workshops – Richard Cyganiak, Arofan Gregory, Wendy Thomas, and Joachim Wackerow – for their support and encouragement in developing the XKOS ideas. The authors also wish to thank the participants not already mentioned in the XKOS development group: Thomas Bosch, Rob Grim, and Jannik Jensen.

Notes

1. Franck Cotton, Institut National de la Statistique et des Études Économiques. Daniel W. Gillman, US Bureau of Labor Statistics, Yves Jaques, Food and Agriculture Organization of the United Nations.

The opinions in this paper are those of the authors only and do not necessarily reflect the policies and programs of the Institut National de la Statistique et des Études Économiques, the US Bureau of Labor Statistics, or the Food and Agriculture Organization of the United Nations.

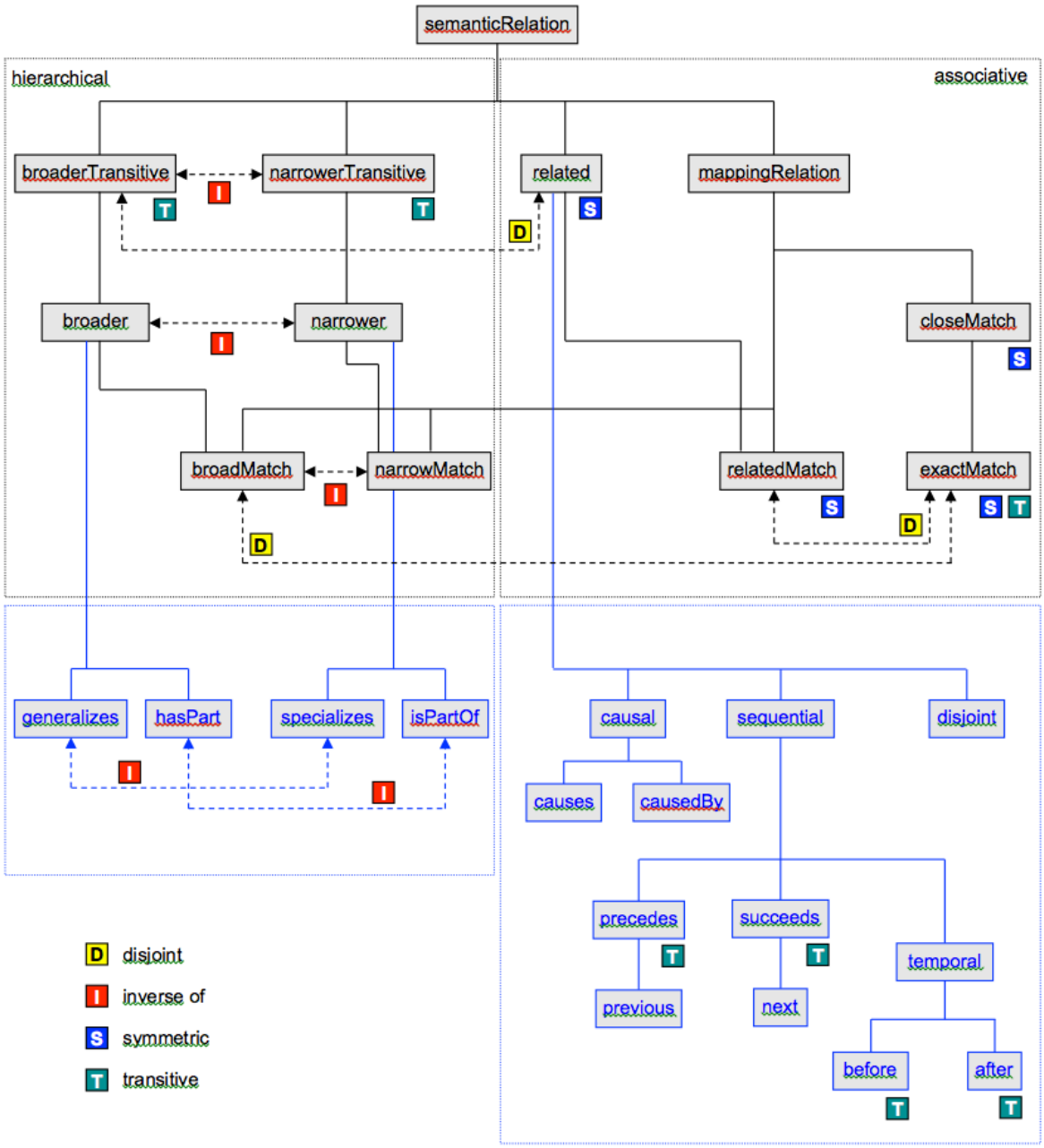
2. http://www.w3.org/2004/02/skos
3. http://www.w3.org
4. http://en.wikipedia.org/wiki/Semantic_Web and http://semanticweb.com/tag/tetherless-world-constellation
5. http://linkeddata.org
6. http://www.w3.org/DesignIssues/LinkedData.html

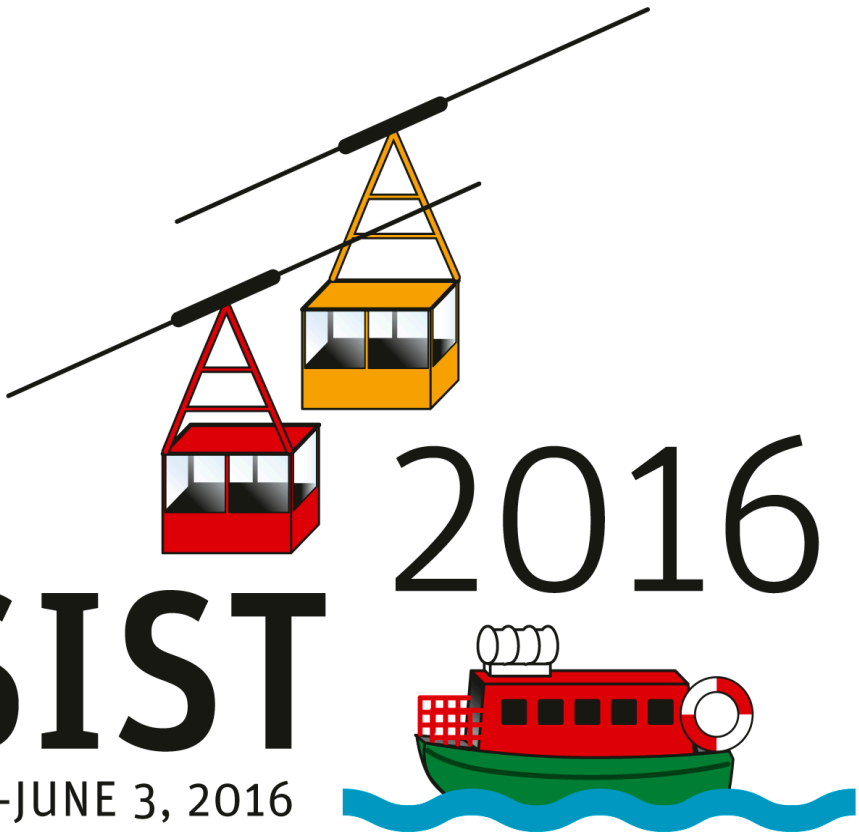
7. http://www.w3.org/RDF
8. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.40/2010/wp.4.e.pdf
9. http://linkeddata.org
10. http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/cube.html
11. http://data.gov.uk
12. http://www.dagstuhl.de
13. http://www.dagstuhl.de/en/program/calendar/evhp/?semnr=11372
14. http://www.dagstuhl.de/en/program/calendar/evhp/?semnr=12422
15. http://www.w3.org/2006/07/SWD/SKOS/reference/20090315/implementation.html
16. http://www.w3.org/TR/rdf-schema/
17. http://tools.ietf.org/html/rfc3986
18. http://www.ietf.org/rfc/rfc1738.txt
19. http://xmlns.com/foaf/spec/
20. http://www.w3.org/TeamSubmission/turtle/
21. http://www.obitko.com/tutorials/ontologies-semantic-web/rdf-graph-and-syntax.html
22. http://www.w3.org/TR/skos-primer/
23. http://www.bls.gov/soc/
24. http://www.bls.gov/iif/oshoiics.htm
25. http://www.bls.gov/tus/lexicons.htm
26. http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=14319930
27. http://www1.unece.org/stat/platform/display/metis/The+Common+Metadata+Framework
28. http://eurovoc.europa.eu/
29. http://id.loc.gov/authorities/subjects.html
30. http://dublincore.org/documents/dcmi-terms/
31. http://www.w3.org/ns/radion
32. http://unstats.un.org/unsd/cr/registry/isic-4.asp
33. http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-07-015/EN/KS-RA-07-015-EN.PDF

Annex 1

SKOS and XKOS properties relating concepts

Note: SKOS properties are in the two upper boxes, XKOS in the two lower.





IASSIST 2016 will take place in Bergen, Norway, hosted by the Norwegian Social Science Data Services.

For any questions - please contact:
heidi.tvedt@nsd.uib.no

IASSIST

INTERNATIONAL ASSOCIATION FOR SOCIAL SCIENCE INFORMATION SERVICE AND TECHNOLOGY

ASSOCIATION INTERNATIONALE POUR LES SERVICES ET TECHNIQUES D'INFORMATION EN SCIENCES SOCIALES

The **International Association for Social Science Information Service and Technology** (**IASSIST**) is an international association of individuals who are engaged in the acquistion, processing, maintenance, and distribution of machine readable text and/or numeric social science data. The membership includes information system specialists, data base librarians or administrators, archivists, researchers, programmers, and managers. Their range of interests encompases hard copy as well as machine readable data

Paid-up members enjoy voting rights benefit from reduced fees for attendance at regional and international conferences sponsored by **IASSIST**. Join today by filling in our online application:

http://www.iaassistdata.info/

Online Application

IASSIST Member ($50.00 (USD))
Subscription period: *1 year, on: July 1st*
Automatic renewal: *no*

Please fill in the information our Online Form

The application is in USD, however, we do accept Canadian Dollars, Euro, and British Pounds as well.

The membership rates in all currencies as well as the Regional Treasurers who manage them are listed on the Treasurers page