XKOS: An SKOS Extension for Statistical Classifications

Franck Cotton[1], Richard Cyganiak[2], R.T.A.M. Grim[3], Daniel W. Gillman[4], Yves Jaques[5], and Wendy Thomas[6,7]
[1]Institut National de la Statistique et des Études Économiques, Paris, FRANCE
[2]DERI, Galway, IRELAND
[3]Tilburg University, Tilburg, NETHERLANDS
[4]Bureau of Labor Statistics, Washington D.C., USA
[5]Food and Agriculture Organization of the United Nations, Rome, ITALY
[6]Minnesota Population Center, Minneapolis, MN, USA
[7]Corresponding author: Wendy Thomas, email: wlt@umn.edu

Abstract

Resource Description Framework (RDF) based approaches to data management are growing in use and acceptance as governments and international organizations increasingly embrace the Semantic Web and Linked Open Data (LOD). This paper describes the efforts to marry LOD techniques to the needs of the international statistical community during two workshops on Semantic Statistics held at Schloß Dagstuhl, Leibniz Center for Informatics with the support of the GESIS Leibniz Institute for Social Sciences and the Data Documentation Initiative (DDI). Specifically, we address the extension of the Simple Knowledge Organization System (SKOS), an RDF vocabulary, to satisfy the requirements of classification systems and concept management for the statistical community.

The SKOS vocabulary was heavily influenced by the needs of thesauri, which rely on loosely defined notions of 'broader than', 'narrower than, and 'related to' relationships. Statistical classifications on the other hand use more formally defined hierarchical relations, which are referred to as generic (generic-specific) and partitive (whole-part). Further, statistical classifications, through their hierarchies, are structured according to levels that correspond to increasingly detailed views of the field covered. Finally, statistical concept management requires the use of associations that are more specific than the generic 'related to'. Causal, sequential, and temporal relations therefore need to be defined. XKOS addresses the greater detail required to manage statistical classifications by extending the existing SKOS definitions of object classes and relationships.

The proposed extensions to SKOS are guided by the needs of the statistical community as well as the requirements laid out in ISO standards on terminology, as reflected in ISO 704:2000 ([ISO704]) and ISO 1087-1:2000 ([ISO1087]). These standards describe and define the constructs and relations necessary for concept management and for a more complete description of statistical classifications. This paper describes concepts, their designating terms and codes, their associative relations, their structures and finally the relationship between concepts and the real-world objects they classify. The goal is to provide a common means for making statistical classifications available through emerging web technologies, while also building on the RDF good practices of reuse.

Key Words: statistical classifications, concepts, RDF, Semantic Web, SKOS

1. Introduction

As statistical agencies increase their use of the web to disseminate and expose their data to the public, their use of Resource Description Framework (RDF) has grown. RDF is a W3C standard designed to facilitate the publication of structured data on the Web. RDF defines a basic structure of Subject-Predicate-Object (known as a "triple") to describe objects and their relationships. It provides a vocabulary for describing a basic set of classes and properties and a means of extending the vocabulary into new domains

through the use of namespaces (RDF, 2004). This allows it to be endlessly extensible and yet continue to maintain the unique identity of objects in a rapidly expanding environment.

The Linked Open Data (LOD) community leverages the RDF framework by providing a technique for organizing data and metadata on the Web in a way that allows users to find, understand and combine data from multiple resources. The use of LOD by libraries, archives, GIS, statistics agencies and others is seen as a means of supporting the growing demand for "open data" and "open access". The concern within the statistical community is to develop a framework within the LOD world that supports the accurate and suitable use of statistics in a consistent way. For example the development of the Data Cube vocabulary is a rendition of the SDMX Information Model for RDF (Data Cube, 2010).

However, the needs of the statistical community do not lie solely with the dissemination of statistical data tables. Providing access to underlying microdata and/or its related metadata as well as being able to clearly manage concepts and classification systems are two related areas of need within the statistical community. During two workshops on Semantic Statistics held at Schloß Dagstuhl, Leibniz Center for Informatics with the support of the GESIS Leibniz Institute for Social Sciences and the Data Documentation Initiative (DDI), work was begun to address these needs by bringing together LOD experts and members of the international statistical community. This paper focuses on the work done during and following these two workshops to address statistical concept and classification management. Specifically, we describe the extension of the Simple Knowledge Organization System (SKOS), an RDF vocabulary dedicated to the representation of concept schemes, to satisfy the requirements of classification systems and concept management for the statistical community.

2. Descriptive Needs of Statistical Classification and Content Management

The requirements for statistical classification management are laid out in ISO standards on terminology, as reflected in ISO 704:2000 (revised 2009) ([ISO704]) and ISO 1087-1:2000 ([ISO1087]). These standards describe and define the constructs and relations necessary for concept management and for a more complete description of statistical classifications. Typical thesauri, taxonomies, subject heading systems, and classification schemes contain standard hierarchical and associative relationships. The hierarchical concepts of broader and narrower transitive relationships (broader, narrower, broad match, narrow match) and associative relationships of related and related match, as well as mapping relationships (close match, related match, and exact match) are generally covered. These are the types of common relationships and usages that are provided within Simple Knowledge Organization System (SKOS) (SKOS, 2012).

Statistical classifications on the other hand use more formally defined hierarchical relations, which are referred to as generic (generic-specific) and partitive (whole-part). Moreover, statistical classifications, through their hierarchies, are structured according to levels. Levels correspond to all concepts found at the same distance from the top of a hierarchy and are used to provide increasingly detailed views of the field covered by the classification. Finally, statistical concept management requires the use of associations that are more specific than the generic 'related to': causal, sequential, and temporal relations need to be defined. The eXtended Knowledge Organization System (XKOS) is being developed to address the greater detail required to manage statistical classifications by extending the existing SKOS definitions of objects and relationships. The goal is to

provide a common means for making statistical classifications available through emerging web technologies, while also building on the RDF good practices of reuse.

3. SKOS Coverage

SKOS focuses on the common requirements of classification and thesauri systems (SKOS, 2009).

- Conceptual relationships – hierarchical relationships of *broader than* (BT), *narrower than* (NT) and the associative relationship *related to* (RT)
- Concept Scheme Extension – the ability to expand a scheme over time where new concepts may refer to existing concepts
- Mapping concepts from different concept schemes
- Representation of labels
- Support for multiple natural languages
- Local specialization – the ability to extend SKOS vocabularies
- Representation of descriptions

SKOS is being used as a basis for extension because it provides a well-known means of representing knowledge organizations systems using RDF. It already supports the basic framework needed for statistical classification, since it encompasses the idea of a Concept Scheme, a structured representation of a set of concepts or categories and their relationships that can be managed over time. The concepts can be described and labeled using associated names or codes. Semantic relationships between concepts (between parents and children) can be expressed. In addition, basic associations can be made (related, close match, exact match). In short, SKOS provides the underlying framework for statistical classification but lacks the refinements needed by the statistical community.

XKOS is not the first extension of SKOS. Other SKOS extensions exist, for example SKOS-XL, designed to support needs of the multi-lingual thesaurus community with the use of labels as class instances rather than simple literals. In other words, allowing labels to take part in relationships and have their own properties. Using the existing SKOS features and extending these where needed, will allow the statistical classifications described with the XKOS specification to be found using the same tools which identify and provide access to other SKOS based content on the Web.

4. XKOS

XKOS also borrows from the Neuchâtel Model (Neuchâtel, 2009) but is not a complete translation of this model. For example, the classification indexes are not supported. The development of XKOS focuses on the following functional areas:

- Structural definition
- Specification of textual properties
- Refinement of the semantic properties (hierarchical and associative)
- Correspondence between classification schemes

Each of these functional areas will be discussed indicating which SKOS classes were used and how and why they have been extended using XKOS. Further details may be found in *eXtended Knowledge Organization System (XKOS)* (Cotton, *et al,* 2013).

4.1 Structural Definition

Whereas SKOS contains the ideas of a concept (*skos:Concept*) and classification scheme (*skos:ConceptScheme*), it does not specifically encompass the idea of a classification. A classification is basically a set of classification schemes of the same name. So that the International Standard Industrial Classification (ISIC) would be the *classification* and each major version would be a *classification scheme.* A classification scheme is an aggregation of concepts and their semantic relationships.

XKOS uses SKOS classes and related properties (codes, labels, etc.) to represent the following:

- Classification items (*skos:Concept*)
- Classification Scheme (*skos:ConceptScheme*)
- Classification (*skos:Concept*) which can in turn be part of concept schemes representing classification families

What XKOS adds is a set of properties to describe the links between these objects. For example, a classification scheme may be attached to a classification using *xkos:belongsTo* while *xkos:follows* or its sub-property *xkos:supercedes* defines the relationship between successive versions of a classification. Where possible, XKOS makes use of existing properties in keeping with the RDF spirit of re-use particularly from widely supported vocabularies.

XKOS also defines a generic property, *xkos:classifiedUnder*, that can be used to link a resource to the classification item it belongs to (for example an enterprise to an item in a classification of economic activities). As this may become quite complex in some cases, it is expected that further specialization of *xkos:classifiedUnder* will be made.

While the structure of a classification scheme can be described using SKOS properties to define membership in a scheme (*skos:inScheme* or *skos:member*) as well as basic hierarchical relationship (*skos:broader*, *skos:narrower*), SKOS has no representation of levels of classification. XKOS has defined a subclass of *skos:Collection* to do this. Levels (*xkos:ClassificationLevel*) are structured within an RDF List organized from most aggregated on down. The list can be further defined using the *xkos:depth* property to express the distance of the specified level from the root node. Generic names for given levels (e.g. “section”, “division”, etc.) can be attached to a level using the *xkos:organizedBy* property. The list itself is attached to the classification scheme using the *xkos:levels* property while individual level items are attached to their *xkos:ClassificationLevel* using the *skos:member* relationship.

4.2 Textual Properties

Classification schemes have a number of specified areas of description associated with a classification item that are not well represented in SKOS. SKOS contains, among others, both a generic note (*skos:note*) and scope note (*skos:scopeNote*). XKOS has introduced sub-properties for *skos:scopeNote* that allow greater specification for the type of notes commonly found within classification systems. An *xkos:inclusionNote* provides a definition of what is included within a classification item and can be further refined by *xkos:coreContentNote* and *xkos:additionalContentNote*. The property *xkos:exclusionNote*, a companion to *xkos:inclusionNote*, describes specific exclusions from the scope description of a classification object.

4.3 Semantic Properties

As noted earlier statistical classification requires the ability to specify loosely described broader and narrower hierarchical relationships and to define both generic (generic-specific) and partitive (whole-part) aspects of those relationships. XKOS has provided a set of properties to refine the skos:broader and skos:narrower along these two dimensions.

XKOS refines:	***skos:narrower***	***skos:broader***
Generic dimension	*xkos:generalizes*	*xkos:specializes*
Partitive dimension	*xkos:hasPart*	*xkos:isPartOf*

In additions several refinements have been made to the associative *skos:related* to type the relationship as causal, sequential, or disjoint. Causal relationships are identified using the property *xkos:causal* which may be further refined by *xkos:causes* or *xkos:causedBy*. Sequence is identified with *xkos:sequential* which includes the transitive pair *xkos:precedes* (with sub-property *xkos:previous*) and *xkos:succeeds* (with sub-property *xkos:next*). Temporal relationships are identified with *xkos:temporal* and further refined by *xkos:before* and *xkos:after*. There is also the relationship identified with *xkos:disjoint* which has no further refinements.

4.4 Correspondence

A number of fields have several classification schemes covering the same classifications for different organizations, different purposes, or simply different periods of time (new versions). These semantically related classification schemes need to be mapped, describing the correspondences between classification schemes and tracking how classification items have been created, split, merged, or been removed from active use. This is an area that SKOS does not address but it is vital to classification management within the statistical community.

XKOS defines the *xkos:Correspondence* class which consists of (*xkos:madeOf*) a set of associations (*xkos:ConceptAssociation*). The *xkos:ConceptAssociation* is similar to the Correspondence Item in the Neuchâtel model, but is not limited to a pair-wise comparison. SKOS concepts (*skos:Concept*) are classified as "source" or "target" for the purpose of comparison. A "source" concept is linked to an *xkos:ConceptAssociation* using *xkos:sourceConcept* while a "target" concept is linked to the same *xkos:ConceptAssociation* using the *xkos:targetConcept.* A future version of XKOS may define additional properties or sub-classes for *skos:Correspondence* and *xkos:ConceptAssociation* to further describe different types of correspondence or the typology of item changes detailed in the Neuchâtel model (Annex 3).

5. Conclusions

As the statistical community moves more fully into the world of LOD, providing structures that will clearly integrate the detailed and specific information required for the accurate use of statistical data becomes increasingly important. By extending a widely used specification like SKOS, we provide an entry point for RDF based tools within the LOD community to discover and effectively integrate statistical data sources into the semantic knowledge network. It parallels the approach taken in the bibliographic community where numerous domain specific bibliographic specifications have existed

for hundreds of years. Dublin Core has become the high-level basic common set of descriptors for bibliographic records and allows linking between objects as the high level across a broad set of domains. Each domain retains its refinements because they are important to that domain, however, by expressing themselves as extensions of (or mapping to) Dublin Core they expose their contents to a broader world of knowledge, supporting linkages that were previously difficult or impossible to support.

The intent of XKOS is to provide the detailed descriptions and relationships required by the statistical community for the management of statistical concept and classification systems. Doing this by extending a well-known specification such as SKOS, which is already supported by LOD RDF-based tools, should allow XKOS users to leverage existing tools and approaches for exposing their classification systems more broadly. Its success will be based on the uptake of XKOS by the statistical community. It is hopeful that others who require these types of specializations will also be able to make use of these extensions.

XKOS is still a work in progress. A few unresolved issues remain and we hope that the users of XKOS will provide comments and guidance to the authors on the effectiveness of XKOS as we prepare to submit the standard as a W3C Editor's Draft.

6. 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 wish to thank the participants not already mentioned in the XKOS development group: Thomas Bosh and Jannik Jensen.

References

Cotton, F., Gillman, D.W, and Jaques, Y. (2013). "eXtended Knowledge Organization System (XKOS)", *Joint UNECE/Eurostat/OECD Work Session on Statistical Metadata*

"Neuchâtel Model – Classifications and Variables" (2009), http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=14319930

"The RDF Data Cube vocabulary" (2010), http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/cube.html

"RDF Vocabulary Description Language 1.0: RDF Schema" (2004), http://www.w3.org/TR/rdf-schema/

"SKOS Simple Knowledge Organization System – Home Page" (2012). http://www.w3.org/2004/02/skos/

"SKOS Use Cases and Requirements" (2009) http://www.w3.org/TR/skos-ucr/