XKOS: Extending SKOS for Describing Statistical Classifications

Franck Cotton, Daniel W. Gillman, and Yves Jaques

Institut National de la Statistique et des Études Économiques, Paris, France
`franck.cotton@insee.fr`
US Bureau of Labor Statistics, Washington, USA
`Gillman.Daniel@bls.gov`
UN Population Fund
`jaques@unfpa.org`

Abstract. A statistical classification scheme is used in statistics to place units in one and only one category from that scheme. These categories are used as units of analysis; dimensions in databases, tables, and time series; criteria for subdividing or aggregating populations; etc.

SKOS provides a simple model for rendering a statistical classification scheme in Linked Open Data (LOD) format. However, it is too simple for some purposes, so XKOS (eXtended Knowledge Organization System) was designed to extend it in order to allow richer representations of statistical classifications.

This paper describes the statistical needs to address and the results of the XKOS design work, including some examples on well-known statistical classifications. Concrete implementations of the vocabulary are also provided, with examples of real use cases.

Keywords. SKOS, XKOS, classification, scheme, statistics, linked data.

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. XKOS is an extension of the Simple Knowledge Organization System (SKOS) [1] 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.

The original SKOS is used widely in LOD applications, as seen in the SKOS Implementation Report[1]. As a result, a group was formed at the Dagstuhl Workshops held at Schloβ Dagstuhl[2] in Germany on *Semantic Statistics for Social, Behavioural,*

---

[1] http://www.w3.org/2006/07/SWD/SKOS/reference/20090315/implementation.html
[2] http://www.dagstuhl.de

and Economic Sciences: Leveraging the DDI[3] Model for the Linked Data Web in September 2011[4] and October 2012[5]to look at the suitability of using SKOS in the statistical data community for LOD work [2].

In this paper, we provide introductory remarks to set the stage for discussion, provide a short primer on statistical classifications, describe limitations of SKOS to statistical classifications, 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 through the use of examples. In a last section, we give examples of concrete implementations of the vocabulary.

2   Statistical Classifications

A statistical classification scheme (SCS) is representable as a SKOS Concept Scheme and used for statistical purposes. In the next section, we illustrate the need for extensive semantics to account for the meanings conveyable in SCSs. Here, we say what an SCS is used for by statistical organizations.

An SCS is a hierarchical *skos:ConceptScheme* which includes concepts, associated codes (numeric string labels), short textual names (also labels), definitions, and longer descriptions that include rules for their use. Examples of SCSs used in the US statistical agencies are

- Standard Occupational Classification[6] (SOC)
- North American Standard Industrial Classification[7] (NAICS)

By a hierarchy, we mean a system of concepts where each has zero or one parents and zero or more children. The root, or top concept, has no parent; the leaves, or bottom concepts, have no children; and the rest have one parent and one or more children.

An SCS may be a flat list (i.e., one level) or a hierarchy, and if it is a hierarchy then it is with the added proviso that all its concepts are grouped into levels. In every SCS, the root concept is implicit and is often referred to as the defining concept for the SCS. All the concepts in one level are the same number of relationships away from the root, and this number is known as the depth of the level. All the concepts at each level are mutually exclusive and exhaustive (ME&E), meaning each unit can be classified to one and only one concept per level. For example, the first level in NAICS is known as Sectors, and these sectors are the broadest industry categories.

From the perspective of SKOS, the relation between a concept and its parent is one of the *broader than* and *narrower than* kinds, depending which direction one is look-

---

ing. A parent is a broader concept, whereas a child is narrower. We have more to say about this in the next section.

Each level is defined by its own concept, as the collection of concepts at a particular level have a common overall meaning. For instance, the first level of NAICS below the root (i.e., industry or economic activity) is the sector, and Government is one of the sectors.

The most important use of a classification scheme is to classify and organize units within some domain, for example business establishments by industry. NAICS is used for this in the US, but due to limitations in coverage for some geographic areas, establishment sizes, or NAICS concepts, not enough data might be available to report meaningfully at all levels. So, the lowest level with meaningful data in most of the concepts is used. However, what determines "meaningful" is a statistical consideration, not germane to SKOS, and out of scope for this paper. The interested reader can consult a text on statistical sampling [3].

Practically speaking, the rows in tables[8] and the dimensions used to specify measures in a time series (for instance the US Consumer Price Index for specific expenditures and municipalities[9]) are major uses of SCSs in data dissemination. Tables and time series report aggregated data that are classified by SCSs. Some of the SCSs may be hierarchies, so one level is chosen to report data. If the data are sparse at that level, then there is a danger that data about individuals (people or businesses) is recoverable. In this case, the next higher level is used, and the data are aggregated some more into the broader categories. This is an important reason for the hierarchical design of SCSs.

Another use of SCSs is in data collection. The possible answers to questions on a form or in a questionnaire are the categories in SCSs. When the range of answers covers all that are possible, and since answer choices must be mutually exclusive, the ME&E criterion for SCS is satisfied. Another way SCSs are used in data collection is through classifying textual responses. For instance, the US American Community Survey asks respondents to briefly describe their jobs, and these descriptions are classified to NAICS and SOC. The levels selected are based on the detail the questions are designed to elicit. The US Survey of Occupational Injuries and Illnesses asks respondents to describe incidents in their workplaces that resulted in loss of work time. These incidents are classified into the nature of the injury or illness, the affected body part, the source of the malady, and the event that caused it (see section 3.2).

Statistical agencies that manage SCSs sometimes make versions to reflect changes in the subject matter domain, and these versions are separate SCSs. However, they belong to the same family, and that is known as a classification. For instance, NAICS is updated every 5 years, and each version (a separate SCS) is known by its year.

A model for describing and managing SCSs developed by the international statistical community can be found in [4].

---

[8] http://www.bls.gov/news.release/empsit.a.htm
[9] http://www.bls.gov/cpi/cpid1305.pdf

3   SKOS Limitation

3.1   SKOS and What Is Missing

SKOS provides a means for representing knowledge organization systems using RDF, and this makes the use of SKOS applicable to SCSs. It is beyond the scope of this paper to provide a detailed description of SKOS. We direct the interested reader to the SKOS web site[10]. However, SKOS contains the following basic ideas, whose definitions we paraphrase here:

- Concept Scheme – any knowledge organization system (including SCSs)
- 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 an SCS. We can account for the scheme itself (*concept scheme*), all its underlying concepts with concept, what each concept means (*definition*), the labels and codes associated with a category (*label* / *notation*), and relationships with its parent and all its children (*semantic relation*).

SKOS is based on ISO 25964-1 [5]. This standard describes three basic kinds of hierarchical relations between concepts: *generic*, *partitive*, and *instantiation*.

Interestingly, SKOS provides 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[11], this simplification is acknowledged.

SKOS also specifies an *association* relation between concepts, but this is not made any more detailed. Possible detail might include *sequential*, *temporal*, and *causal* relations. The *sequential* relation refers to ideas where one is the antecedent of the other, either temporally or spatially, such as between production and consumption. The specialized *temporal* relation is based on time, such as between spring and summer. Finally, the *causal* relation relates cause and effect, such as the detonation of a hydrogen bomb and nuclear fall-out. See [6] for further explanation of the relations described here and above.

Because levels have a concept associated with them and they have depth (from the root), there is no satisfactory way to account for them in SKOS. Thus, we need to add the notion of level. It is a kind of *skos:Collection*.

3.2 Examples

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

1. The US Standard Occupational Classification System[12] (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[13] (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.

---

[12] http://www.bls.gov/soc/
[13] http://www.bls.gov/iif/oshoiics.htm

3. The US American Time Use Survey — Activity Coding Lexicons[14], 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
0401 –   Caring For & Helping non-Household Children
040104 –   Arts & Crafts with non-Household Children
040112 –   Dropping Off/Picking Up non-Household Children

<u>Dropping off non-household children</u> is a sequential activity related, in this example, to having <u>supervised arts-and-crafts activities</u> (or some other activity in the 04 group) previously. So, there are associations between some pairs of activities within this classification, though they are not intrinsic to the SCS. 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.

4   XKOS

In the following, we list some of the extensions XKOS contains and guide the interested to reader to another paper that contains more detail[7].

- *xkos:belongsTo* is used to attach a classification scheme to its classification
- *xkos:follows* or its sub-property *xkos:supercedes* is used to relate classification schemes that are successive versions
- *xkos:classifiedUnder* is used to indicate a unit is classified by some concept
- *xkos:ClassificationLevel* (subclass of *skos:Collection*) is the level; the levels of a classification scheme are structured as an RDF List, starting with the most aggregated, and the list attached to the classification scheme by the *xkos:levels* property.
- *xkos:depth* property expresses the distance of a given level from the root node of the hierarchy
- *xkos:organizedBy* property can be used to record the generic name of the items of a given level (e.g. "section", "division", etc.).
- explanatory notes use *xkos:coreContentNote*, *xkos:additionalContentNote*, *xkos:inclusionNote*, and *xkos:exclusionNote*, which are sub-properties of *skos:scopeNote* and correspond to a typology of explanatory notes widely used for statistical classifications
- *xkos:ConceptAssociation* is a class that can be used to represent correspondences between classification items across SCSs through input or source *skos:Concept*(s) and output or target *skos:Concept*(s).
- *xkos:Correspondence* groups a set of *xkos:ConceptAssociation*(s) to represent a concordance or correspondence table between two classification schemes (for example two versions of a classification).

---

[14] http://www.bls.gov/tus/lexicons.htm

- *xkos:specializes* and *xkos:generalizes* represent each side of the *generic* relation.
- *xkos:isPartOf* and *xkos:hasPart* represent each side of the *partitive* relation.
- *xkos:disjoint* property is used to explicitly state that two given concepts do not overlap
- *xkos:causal* is subdivided into the directional *xkos:causes* and *xkos:causedBy*, to express causality
- *xkos:sequential* indicates that two concepts in a scheme are in a sequential relationship
- *xkos:succeeds* and *xkos:precedes* are used when a sequence has a known order and are further refined by *xkos:previous* and *xkos:next*, the immediate successor or predecessor
- *xkos:temporal* is used when a sequence is of a temporal nature and is subdivided into the directional *xkos:before* and *xkos:after*

5   Implementing XKOS

A first example of XKOS utilization can be found on INSEE[15]'s linked data site[16]. The Institute published different statistical code lists and classifications there, and notably the NAF, the French refinement of the European NACE, is one.

This publication makes use of different XKOS features, for example classification levels, maximum-length labels and explanatory notes. This allows making useful requests which would be more difficult, or even impossible, using only a SKOS representation. As illustration, the following SPARQL query gives the list of all the NAF divisions, which is a very common use case for classifications.

```
PREFIX skos:<http://www.w3.org/2004/02/skos/core#>
PREFIX xkos:<http://purl.org/linked-data/xkos#>

SELECT ?code ?label WHERE {
  ?item skos:notation ?code .
  ?item skos:prefLabel ?label .
  ?item skos:inScheme
<http://id.insee.fr/codes/nafr2/naf> .
  ?level skos:member ?item .
  ?level xkos:organizedBy
<http://id.insee.fr/concepts/nafr2/division> .
  FILTER langMatches(lang(?label), 'en')
} ORDER BY ?code
```

A second example was created especially for this paper and is based on the international classifications published by the United Nations Statistics Division[17]. We chose

---

[15] INSEE, Institut National de la Statistique et des Études Économiques, is the French National Statistical Institute.

[16] http://rdf.insee.fr/

the ISIC (International Standard Industrial Classification), which plays a central role in the international system of economic classifications[18]. The objective was to represent in XKOS the last two revisions of the ISIC and the historical correspondences between them.

The main challenge for this operation was to be able to analyze the explanatory notes and to split them into the specific categories defined by XKOS (core or additional content, exclusions, etc.). This can be done by recognizing patterns in the note text ("This group contains", "This division excludes", etc.), but, although the ISIC notes are usually well structured, there are variations of these patterns and numerous special cases, so that the process cannot be fully automated.For example[19], the note for division 23 combines in a single paragraph the descriptions of the core and additional contents for this division.

The ISIC is available in different formats on the UNSD web site, but for the explanatory notes, the only possibilities are PDF, MS Access, and the online HTML publication. Correspondences between ISIC revisions or with other classifications are also available as HTML, or in downloadable plain text files. For the sake of simplicity and coherence, we decided to use the HTML online publication as the reference source of data.

As a consequence, a simple Java application was developed in order to retrieve the data from the web site, based on the Jericho HTML parser[20]. Though ad hoc, this application was designed to be modular and should be easily adaptable to other use cases.

In a first step, which is repeated for ISIC revisions 3.1 and 4, the web pages describing the classification items are requested recursively over HTTP, starting from the page describing the top structure, and the relevant information about classification hierarchy, codes, labels, and notes is extracted and copied in a simple XML file. This first step is also the occasion to make an automated categorization of the notes based on regular expressions.

A paragraph starting with "This class also contains:" will be considered an XKOS note of type "additional content". Another starting with "This section includes" will be a "core content note", etc.Further pattern matching is then performed to verify that two types of notes are not grouped in one paragraph (in the first example, we will check if the text contains "exclude", in the second we will also search if it contain "also"). If any of these additional matches is positive, a log entry will be recorded and manual verification will ensue.

A paragraph that does not match any predefined expressions will take the type of its predecessor if it has one, or be categorized as "general". Here again, the assumption will be logged for further human checking, because some exceptions may occur

---

[17] http://unstats.un.org/unsd/cr/registry/regct.asp?Lg=1

[18] This system is described for example in the official NACE Rev.2 publication at http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-07-015/EN/KS-RA-07-015-EN.PDF, pp 13-14.

[19] All reference material for the example can be accessed through the ISIC main page at http://unstats.un.org/unsd/cr/registry/regcst.asp?Cl=27&Lg=1

[20] http://jericho.htmlparser.net

(see for example Section B, where general notes can be found between the descriptions of the included and excluded contents).

The code below is an extract of the XML file produced by the Java program, showing a simple example of note analysis:

```
<Item code="1430" parent="143">
  <Label lang="en">Manufacture of knitted and crocheted
apparel</Label>
  <Notes lang="en">
    <Sequence>C X</Sequence>
    <Structure>0:C 3:X</Structure>
    <Elements>
      <Element><div>This class includes:</div></Element>
      <Element><div class='item'>-  manufacture of knit-
ted or crocheted wearing apparel and other made-up ar-
ticles directly into shape: pullovers, cardigans, jer-
seys, waistcoats and similar articles</div></Element>
      <Element><div class='item'>-  manufacture of ho-
siery, including socks, tights and panty-
hose</div></Element>
      <Element><div>This class excludes:</div></Element>
      <Element><div class='item'>-  manufacture of knit-
ted and crocheted fabrics, see 1391</div></Element>
    </Elements>
  </Notes>
</Item>
```

The HTML code extracted from the web pages is placed in the `<Elements>`node. In the ISIC case, and more generally for classifications published by the UNSD, explanatory notes are organized in `<div>` tags with a `class` attribute indicating the list level, but other organizations may use different conventions.

The `<Sequence>` element gives an overall vision of the notes structure as determined by the program: in this case a "core content" (`C`) note is followed by an "exclusion note" (`X`). The `<Sequence>` element values can easily be queried by XPath to detect anomalous structures. The `<Structure>` element completes the `<Sequence>`information by giving the start index of each part in the `<Element>` list: this will be used by the following processing step.

The `<Structure>` and `<Sequence>` elements can be edited in the XML file to correct possible errors made by the program. As explained above, this step can be guided by the detection of unusual sequences and bythe warnings logged by the program. For this first experimentation, the corrections were made manually with an XML editor, but it is easy to see how a simple GUI could be developed to improve this step.

Once the XML files(one for ISIC Rev.4 and one for ISIC Rev.3.1) are correct, the last step is to transform them intoXKOS representations of the two classification schemes. XSLT transformations are the tool of choice here, and it is relatively

straightforward to design the general structure of a transformation that produces an RDF/XML serialization of the expected result.

A much more difficult question, though, is how to transform the HTML code used in the notes into a RDF literal: should we take only the plain text content, or should we keep in whole or in part the note structure as it is represented in HTML?

See for example the explanatory notes for ISIC Rev.4 class 1030[21]: the description of the core content is organized in embedded unordered lists, and this structure is important to understanding the description. The exclusions are shown in italics (this can probably be left out), and include pointers to other classes ("… see 1061") which are not rendered as HTML links but could be desirable to capture in automated processing oriented formats like RDF.

For the time being, we chose simplicity and put only plain text in the explanatory notes. INSEE's publication of the NAF usesa more refined solution developed for the EuroVoc thesaurus [8], where the notes are typed as `rdf:XMLLiteral`[22] and contain XHTML+RDFa fragments, with links to other concepts represented through a sub-property of `dcterms:references`[23] (see [9] for a more complete description).

The structure of explanatory notes and the links that they contain are a very important feature for statisticians, and this question of how to represent themin LOD formats is clearly one of the areas where common good practice should be developed in the statistical community.

Once the XML files corresponding to the two ISIC versions are produced, they can be uploaded in a Sesame RDF triple store, completed by the information on correspondences which is directly parsed on the UNSD web site with Jericho. The data is then ready to be queried.Examples of interesting queries are:

```
PREFIX skos:<http://www.w3.org/2004/02/skos/core#>
PREFIX xkos:<http://purl.org/linked-data/xkos#>

SELECT ?code ?label WHERE {
  ?class skos:inScheme
<http://unstats.un.org/codes/isic/4/cs> .
  ?class skos:notation ?code .
  ?class xkos:coreContentNote ?note .
  FILTER regex(?note, "wholesale of office furniture")
}
```

This query searches ISIC Rev.4 for "wholesale of office furniture" in a core content note. It returns only the class that includes this activity (4659), whereas the equivalent query would return two answers mixing inclusions and exclusions (4669) if the notes were only generic SKOS scope notes.

---

[21] http://unstats.un.org/unsd/cr/registry/regcs.asp?Cl=27&Lg=1&Co=1030
[22] http://www.w3.org/TR/rdf-concepts/#section-XMLLiteral
[23] See an example at http://id.insee.fr/codes/nafr2/sousClasse/27.11Z/noteExclusions

```
PREFIX skos:<http://www.w3.org/2004/02/skos/core#>
PREFIX xkos:<http://purl.org/linked-data/xkos#>

SELECT ?source ?target WHERE {
  ? source skos:inScheme
<http://unstats.un.org/codes/isic/3.1/cs> .
  ? target skos:inScheme
<http://unstats.un.org/codes/isic/4/cs> .
  ?association xkos:sourceConcept ?source .
  ?association xkos:targetConcept ?target .
  ?association skos:note ?note .
  FILTER regex(?note, "Repair of weapons")
}
```

This query searches the correspondences between Rev.3.1 and Rev.4 for what happened to the repair of weapons activity (the answer is: it moved from class 2927 to class 3311). This is an example of query on notes concerning classification concordances, which is very useful for statisticians, but would not be easily doablewithout the XKOS extensions.

The figure below summarizes the overall logic of this example.
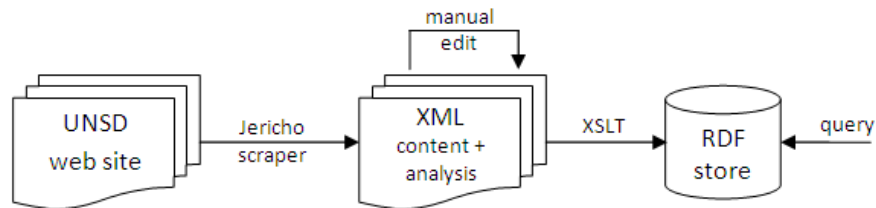
It should be noted that once the data is in XKOS format, it is easy to produce output formats like HTML or PDF.

6   Conclusions

We explained in this paper the rationale for XKOS and described how a simple process could be developed to transform existing information on statistical classifications in XKOS in order to be able to use this information in improved ways. This process can easily be adapted to other situations.

The paper contains a description of a Statistical Classification Systems (SCS), important ways they are used, limitations of SKOS in its ability to describe an SCS, and the several ways we thought SKOS should be extended. From the exposition, it should be clear that the extensions account for the limitations. It is SKOS and the extensions that we call XKOS.

Particular emphasis was placed on the ability to convey semantics, so we were careful to add significantly more relations to XKOS. The examples in statistical offices make clear the need for the additional semantics.

The work to define XKOS is not completed however. Identifying new relations is a priority as well as building a typology of them. The biggest hurdle is to persuade the statistical agencies to use XKOS and build a base of applications as the statistical offices around the world move to adopt Linked Open Data principles.

Finally, we note that although XKOS was developed with the purpose of representing statistical classifications, some elements of the vocabulary can be used outside of the context of classifications. These new applications need further exploration and possible refinement of XKOS.

Acknowledgements

References

1. Isaac, A., Summers, E.: SKOS Simple Knowledge Organization System Primer. Working Group Note, W3C (2009), `http://www.w3.org/TR/skos-primer/`
2. XKOS – DDI/RDF Vocabularies, Downloaded from the Web on 19 July 2013 at `http://www.ddialliance.org/Specification/RDF#xkos`
3. Sharon Lohr, Sampling: Design and Analysis, Brooks/Cole (1999)
4. Neuchâtel Terminology Model - Classification database object types and their attributes, version 2.1, Dowloaded from the web on 19 July 2013 at `http://www1.unece.org/stat/platform/download/attachments/1431993 0/Part+I+Neuchatel_version+2_1.pdf?version=1&modificationDate=12 65695896952`
5. ISO 25964-1 - *Thesauri and interoperability with other Vocabularies*, Part 1: *Thesauri for information retrieval*, ISO (2011)
6. ISO 1087-1:2000 – *Terminology work - Vocabulary*, Part 1: *Theory and application*, ISO (2000)
7. Cotton , F., Gillman, D., and Jaques, Y. (2013) XKOS - An RDF Vocabulary for Describing Statistical Classifications, IASSIST Quarterly (to appear)
8. EuroVoc, the EU's multilingual thesaurus,Downloaded from the Web on 19 July 2013 at `http://eurovoc.europa.eu/drupal/`
9. De Smedt, J., Vatant, B.: The EUROVOC Thesaurus Ontology Schema, Downloaded from the Web on 19 July 2013 at`http://lists.w3.org/Archives/Public/public-esw-thes/2010Feb/att-0023/Ontology.html`