Lexical Conventions

Anita Liang, in collaboration with GI-FAOTERM, Terminology, GICM, FAO, 2005

Lexical conventions create consistent formats for enhancing the effectiveness and efficiency of terminology-related activities. They improve the interpretability of lexical data by humans, serving as a form of documentation (e.g., allowing quick identification of names), and facilitate machine processing (e.g., enabling programmatic manipulations of terms).

The current state of AGROVOC is the immediate motivation for the development of these conventions as it suffers from discontinuities in its development, including a history of changing managers, differences in the conventions (if any) used in the maintenance and extension of the thesaurus during that time, and the application of standards (e.g. ISO 5964 for multilingual thesauri) that in and of themselves do not foster consistent formats. The result is inconsistencies throughout the thesaurus, e.g., in the use of case and singular/plural forms, both within and across languages, which have, in turn, led to other unwanted consequences, e.g., inaccurate translations and classifications.

Apart from the goal of "cleaning up" the terms in the thesaurus, there is also the need to establish lexical conventions for ontology development in anticipation of the medium-term objective to convert AGROVOC into an ontology and of future projects for developing other ontologies within the auspices of the AOS project. The conventions specify consistent labelling and identification of concepts, relations, and attributes while adhering to the rules for legal XML names

Thus, two sets of conventions are specified: one applicable to vocabularies currently being maintained at FAO, that are used mainly (though not exclusively) by humans and that can be applied immediately, and the other to machine-processible semantic structures such as ontologies.

For more information, cf. Soergel *et al*. (2004), Soergel Consultant's report, "Top Quadrant's Guidelines for Ontology Modeling," "Propositions [sic] of Conventions for RDF," *etc*.

FAO vocabularies (dictionary citation form)

General conventions

The conventions for these vocabularies are based on those governing the citation forms of terms found in the dictionary.

Case
All common nouns should be in lower case, e.g., *aphid*, *rhinoceros beetle*. All non-function words (i.e., substantives, verbs, adjectives) occurring in proper noun phrases should begin with an upper case letter, e.g., *Plato, Food and Agriculture Organization, Big Mac,* except if they occur as the initial word of the phrase, e.g., *Of Human Bondage*. Phrase-initial *the* should be dropped. Acronyms should generally occur in upper-case, e.g., *WWW*, except when they have been lexicalized as words in their own right, e.g., *radar*.

Number
All nouns should be in the singular with the exception of such cases as false plurals, e.g., *telecommunications*, proper nouns occurring in the plural, e.g., *Pacific Islands*, and special terminologies, e.g., *Chordates*. All relation names should occur in the singular, e.g., *daughter of, grows in*.

Hyphenation
A variety of rules exist concerning hyphenation (e.g., with the exception of words containing prefixes such as *ex-*, *self-*, and *all-*, most prefixed words should not occur with hyphens) but a general recommendation is to consult a dictionary when in doubt.

FAO ontologies (XML compliant)

The conventions specified above can and should apply to the lexicalized forms occurring in ontologies for the purposes of display (or human interpretation). However, because ontologies are meant to be machine interpretable, other considerations, in addition to those already mentioned, come into play when establishing lexical conventions. For one thing, they must comply with legal XML syntax. For another, ontologies make distinctions beyond the terms and relations characteristic of thesauri, such as property types, instances and classes, and so on. Consequently, separate conventions are necessary for ontology development.

General conventions

Case
Despite the widespread (and sometimes recommended) use of the InterCap style, where all component words are concatenated with no intervening spaces, and each word-initial letter is capitalized (e.g., *MilkProduct, growsIn*), this style will be discouraged. Instead, the dictionary citation form is recommended here, with spaces being replaced by underscores, e.g., *milk_product*. The reason for this is (1) the latter style accommodates other languages (e.g., Chinese, Arabic) whereas the InterCap style is specific to European alphabetic languages and (2) the correct forms of the words making up a given term can be more easily recovered for later processing.

Number
See the rule for thesauri.

Special characters
Remove hyphens, e.g., p*ost-Modernism* should be rewritten as *postModernism*. Replace *&* and */* with *And*. For all other non-alphanumeric characters, replace with the html code for the character (stripped of non-alphanumeric characters), and delimit with periods, e.g., *'* becomes *apos*, and *Sam's* becomes *Sam.apos.s*.

Properties
Properties link two concepts together. We distinguish two types of properties in our ontologies: data properties and object properties.

A data property is a concept whose value consists of a literal. The name of a data property should occur in the singular, e.g., *colour, age, temperature*.

An object property is a concept whose value is another concept. One awkwardness that arises in attempting to name object properties using suggested conventions is the use of the present

tense form of a verb, e.g., *grows_in*, *eats*. The problem is that the truth of the resulting proposition, e.g., may not necessarily be valid in the present, but in the past or for a given time period. Modality (indicating e.g., possibility, necessity) and verbal aspect (the distribution of the action across time) are also not expressed using the present tense convention. As a start to circumventing these problems, a property is added to the relation to create the possibility of specifying temporality (could be enumerated values, past, present, future, or more refined), e.g., *PlantB afflictedBy—[past] PestB*. By default, the value should be present.

When naming static properties, auxiliary verbs like *is* or *has*, should be omitted, e.g., *partOf*, *ingredient*.