GACS: Recommendations for improvements to the thesauri

Thomas Baker and Osma Suominen

Version 1.0 June 11, 2014

1. Technical quality issues

The SKOS versions of the three thesauri were analyzed using the qSKOS[1] vocabulary quality analysis tool to determine whether there are obvious problems in the way the vocabularies are modeled and represented as SKOS. For more detailed descriptions on the quality issues discussed below, see the qSKOS Quality Issues[2] and (Suominen and Mader 2014). Some of the following issues reflect structural problems in the content and structure of the thesauri.

The full lists of affected concepts can be found in the qSKOS analysis reports for each thesaurus, which have been provided to the partner organizations in the Github repository subdirectory qskos-reports[3]. Most of the issues issues highlighted in the report are not on the critical path to creating GACS, but the partners may want to address them anyway in order to improve the quality of their respective thesauri.

1.1. AGROVOC

- **Syntax**: AGROVOC LOD is only available in the TriX syntax, which is not considered a mainstream syntax and is not supported by all RDF toolkits. Apache Jena, for example, requires an extension module. It can be difficult to extract just a single graph from a large TriX file.
- **Empty Labels**: AGROVOC has 220 labels (all of them non-preferred, most with Italian language tag) with empty content.
- **Missing Labels**: AGROVOC has two instances with no label: the concept scheme itself, and a concept with the non-preferred term “forest vitality”.
- **Overlapping Labels**: AGROVOC has 1,039 cases where the same preferred term is used to label several concepts. (Note: the qSKOS report lists 1,886 cases of label overlap, but some of these involve non-preferred labels, for which overlap is less of a problem.)
- **Hierarchical Redundancy**: AGROVOC has 67 cases where a concept has a direct broader link to another concept which is also its ancestor through a longer path. For example, *Age* has the broader concepts *biological properties* and *properties*; the latter is redundant because *biological properties* has a broader link to *properties*. In these cases the redundant relationship could be removed, decreasing the amount of polyhierarchy.
- **Relation Clashes**: AGROVOC has one case in which the SKOS integrity condition S27 is violated, i.e., an associative (`skos:related`) link exists between a concept and one of its ancestors in the hierarchy. These are probably not serious issues (see Suominen and Mader 2014 for discussion), but could indicate modelling issues in the thesaurus as usually a direct or indirect hierarchical relationship should suffice.
- **Disjoint Labels Violation**: AGROVOC has 2,421 cases where the same term is used both as a preferred and as a non-preferred label of the same concept, violating SKOS integrity condition S13.

[1] https://github.com/cmader/qSKOS/

[2] https://github.com/cmader/qSKOS/wiki/Quality-Issues

[3] https://github.com/tombaker/gacswg/tree/master/qskos-reports

1.2. CABT

- **Hierarchical Redundancy**: CABT has 48 cases where a concept has a direct broader link to another concept which is also its ancestor through a longer path.
- **Relation Clashes**: CABT has 70 cases in which the SKOS integrity condition S27 is violated, i.e., an associative (`skos:related`) link exists between a concept and one of its ancestors in the hierarchy.
- **Undefined SKOS Resouces**: The original CABT SKOS dump coined new properties in the SKOS namespace instead of defining a custom namespace URI for them.

1.3. NALT

- **Syntax**: NALT uses an invalid SKOS namespace in the RDF/XML dump, so SKOS-specific tools that expect the standard namespace will not work out of the box.
- **Hierarchical Redundancy**: NALT has 34 cases where a concept has a direct broader link to another concept which is also its ancestor through a longer path.
- **Relation Clashes**: NALT has 24 cases in which the SKOS integrity condition S27 is violated, i.e., an associative (`skos:related`) link exists between a concept and one of its ancestors in the hierarchy.

2. URI policy

A URI policy articulates the method by which URIs are coined using *namespace URIs* and outlines the commitment that an institutional owner has made to their long-term maintenance and persistence. A policy for persistent URIs is required for all thesauri included in GACS.

- **AGROVOC**. The AGROVOC URIs have been available for a relatively long time and are suitable for use in GACS.
- **CABT**. A URI policy is needed for CABT. By leveraging existing identifiers such as the MultiTes ID for the preferred term, for example, opaque, language-independent URIs could be coined in the `cabi.org` domain along the lines of `http://id.cabi.org/cabt/Cnnnn`, where the number `nnnn` is based on the MultiTes ID. Guidance on URI policy is available in the Linked Data book[4], and more specifically for SKOS in Pete Johnston's series of blog posts about creating SKOS from term-based thesauri, particularly Part 2[5] and Part 3[6].
- **NALT**. As with AGROVOC, the NALT URIs have existed for some time and can be used in GACS.

[4] http://linkeddatabook.com/editions/1.0/#htoc10

[5] http://efoundations.typepad.com/efoundations/2011/03/term-based-thesauri-and-skos-part-2-linked-data.html

[6] http://efoundations.typepad.com/efoundations/2011/03/term-based-thesauri-and-skos-part-3-change-over-time-i.html

3. Recommendations

The partner organizations are requested to make these improvements and corrections for their respective thesauri in order to facilitate the creation of GACS. In a strict sense, “consider” points are not required for GACS.

3.1. AGROVOC

- Review the use of uppercase and title case in English, German, Spanish, Portuguese, Turkish, Italian, Polish, and French labels.
- Review the structure of the SKOS representation from the standpoint of complexity and maintainability, especially with regard to the use of relation properties from Agrontology.
- Consider adding (inferred) `skos:narrower` relationships to the published SKOS.
- Consider publishing the graphs included in AGROVOC LOD separately as RDF/XML and/or Turtle files.

3.2. CABT

- Perfect the SKOS representation (non-existent “SKOS” properties, confusion between terms and concepts, URIs based on term strings, chemical and enzyme codes as concepts instead of notations).
- Formulate an explicit URI policy.
- Review the hierarchical structure, particularly the links that cross hierarchies intended to be separate. For example, the BT link from *cereals* to *plants*, illustrated in Figure 1 of the Status Quo report, looks suspicious as it crosses two hierarchies.
- Define a custom vocabulary of SKOS extensions used in CABT, similar to Agrontology but much smaller, for use in the SKOS version.
- Add concept-level provenance data to the SKOS version.
- Consider adding USE-OR information, e.g., as duplicate altLabels, to the SKOS version.
- Consider expressing the thesaurus as SKOS XL. If this is done, consider including term-level provenance information and USE-AND information using the ISO 25964 SKOS extensions[7].
- Define an open license and persistent URI policy, and publish the thesaurus as Linked Data.

3.3. NALT

- Fix the invalid SKOS namespace.
- Add concept-level provenance data to SKOS version.
- Consider expressing the thesaurus as SKOS XL. If this is done, consider including term-level provenance information, and USE-AND information using the ISO 25964 SKOS extensions.

[7] http://purl.org/iso25964/skos-thes#

4. Copyright issues

Each organization providing data for GACS must have a clear policy on data ownership and terms of use, including copyright licensing and related rights such as database rights.

- **AGROVOC**: Define a standard open license (Creative Commons) and apply it consistently. The copyright license for AGROVOC is currently asserted to be both Creative Commons 3.0 Attribution[8] (in a VoiD file[9]) and Attribution-NonCommercial-ShareAlike 3.0 Unported[10] (on the AIMS website[11]).
- **CABT**: CABT is currently not available as Linked Open Data. If its concepts are not identified with URIs and those URIs are not easily accessible on the Web, the mappings between CABT and GACS will be of limited usefulness in open data environments. If CABT, or at any rate the parts of CABT related to GACS, are made available on the Web, the permissible use of the thesaurus data should be clarified by a standard open license (Creative Commons).
- **NALT**: The Terms and Conditions of Use[12] for NALT specify that no license is required to obtain NALT data and specify ad-hoc conditions for non-personal use. Potential users of NALT would find it easier to use the thesaurus if it were made available in terms of a standard Creative Commons license instead of a custom license.

References

Suominen and Mader 2014 Osma Suominen and Christian Mader: Assessing and Improving the Quality of SKOS Vocabularies. Journal on Data Semantics, Volume 3, Issue 1 (March 2014), pp 47-73. Preprint freely available[13]

[8] http://creativecommons.org/licenses/by/3.0/
[9] http://aims.fao.org/aos/agrovoc/void.ttl
[10] http://creativecommons.org/licenses/by-nc-sa/3.0/
[11] http://aims.fao.org/standards/agrovoc/functionalities/download
[12] http://www.nal.usda.gov/web-policies-and-important-links
[13] http://www.seco.tkk.fi/publications/2013/suominen-mader-skosquality.pdf