Guidelines on how to create effective thesaurus concepts

Martin Doerr, Maria Daskalaki, Lida Harami, FORTH-ICS, Heraklion Crete

December 2015

Introduction

The aim of this document is to provide guidelines on how to create concepts for thesauri used as indexing language[1] that should facilitate successful searches in manually indexed information systems of documents or data records, with the preference to reveal as many of the items as possible that are relevant to the research questions posed by domain experts, rather than to exclude possibly unrelated documents. Therefore, it aims neither at the discrimination of scientific concepts against each other nor at identifying terms found in natural language documents that are characteristic for a subject.

The proposed guidelines included in this document are the outcome of an effort by a team of experts[2] of designing and building effective and efficient classification systems for research infrastructures in the humanities[3], based on a consistent methodology which could ensure the intersubjective and interdisciplinary character of its implementation without forcing the experts to abandon their own terminology. Our work focuses on identifying the top-level-concepts (facets and hierarchies) that will become a common basis for thesaurus building, meeting the demands for objectivity and interdisciplinarity. The methodology we propose is based on the principle of faceted classification and the idea that a limited number of top-level concepts can become a substantial tool to harmonize the numerous discipline and even project specific terminologies into a coherent and effective federation (Kramer, 1997) in which consistency can progressively be imposed from the upper layers to

---

[1] As Soergel states: "A **thesaurus** is a structure that manages the complexities of terminology in language and provides conceptual relationships, ideally through an embedded classification/ontology. A thesaurus may specify descriptors authorized for indexing and searching. These descriptors then form a **controlled vocabulary (authority list, index language)**" (Soergel 1998, pp. 16).

[2] Martin Doerr, Maria Daskalaki, Lida Charami, Chryssoula Bekiari, Helen Katsiadaki, Helen Goulis, Makis Chrisovitsanos, Georgia Papadopoulou, Iraklitos Souyioultzoglou, Hella Hollander, Vanessa Hannesschläger, Wolfgang Schmidle.

[3] Dariah EU, VCC3. Thesaurus Maintenance Working Group: A model for sustainable interoperable thesauri maintenance. Draft. To be published.

the lower ones[4]. The method foresees the development of concepts from existing scientific terms by a process of abstraction according to a bottom-up approach. It allows us to exploit all the advantages offered by categorical semantics and to detect the intensional and potential properties of the general concepts under which we can subsume more specific terms. The method has been verified by the construction of the initial "back bone thesaurus" for DARIAH[5].

This document includes of a quick reference guide to the principles, the theoretical presuppositions and to the necessary steps for constructing and building the general concepts that should be used in thesauri, a list of simple principles of "good" and "bad" terms, and an appendix with examples for every step. We put examples in the appendix in order not to break up the simplicity and clarity of the principles.

Terminology

To better understand the guidelines for building an effective thesaurus, we initially provide explanatory notes on the terms we use in order to describe the recommended steps of building the thesaurus. These are:

> ***Source terms***: the terminology of each scientific field, the finite set of general concepts which are used by experts in order to describe their scientific methods, results, tools etc., and which we use as empirical material in order to develop broader terms for them. Source terms are often context dependent. In organizing and building thesauri we regard only source terms which are universals. Instances, being the specific realizations of a general term, such as a placename or a person, are outside the scope of this guideline. They are subject to different methodologies and data structures.
>
> ***Target terms***: the broader terms and top-level concepts which we aim to develop following this guideline. Target terms express types of subjects of attribution i.e. universals whose properties reveal the intensionality (see

[4] See e.g., UMLS (Unified Madical Language System). http://semanticnetwork.nlm.nih.gov/Download/
[5] Dariah EU, VCC3. To be published.

above) of the source terms which are subsumed under the target terms and should be context independent.

***Extension*:** The extension of a term is defined as the set of items for which it is true. It denotes the reference of a term, the range of its applicability by naming the particular items. So the extension of the term 'cat ' is the set of all the cats in the world; the extension of 'red' is the set of all the red things. However, if we define the terms according to their extension, we would not be able to define something that we do not already know or does not already exist. In order to express the meaning of a term we have to refer to its intensionality.

Intension: roughly speaking the intension of a term is the sum of its properties, state of affairs, qualities etc. that constitute the necessary and sufficient conditions for being in the extension of a term/concept. In other words, it is the content of a term, its meaning.

For example, the intension of "bachelor" might be something like: adult, unmarried male. Being an adult, being unmarried, and being male are all necessary conditions for being a bachelor, and their conjunction is a sufficient condition.

However, the fact that the necessary and sufficient conditions of many concepts/terms cannot be detected only through the logical or analytical decomposition of their constituents lead us to additionally introduce the term of *conventional intension* of terms.

Conventional intension of a concept/term consists of properties, state of affairs etc. which are commonly understood and accepted as denoting items belonging to the same extension. Conventional intensions are not merely the result of an (arbitrary) agreement between subjects. On the contrary, the intersubjective agreement on the conventional intensions is based on their reference to a known reality (Millikan, 2010), which exhibits some distinct forms, than a logical determination. For instance, “Human being” is

sufficiently known to us and distinct from other things, even without DNA analysis. For deciding about the criteria upon which we commonly agree to use in order to define the properties of the concepts (intension) with a possible extension, the bottom-up method turned to be a very important methodological tool.

Basic principles and preconditions:

Keystone of our research on concepts' analysis is the fundamental philosophical position that each concept has a purpose or utility. In building thesauri this utility is specified as the ability to deduce from recognizing an item as instance of a concept potential properties of the same item.

In designing and building thesauri we take into account the following rules and preconditions:

1. The definitions of the target terms should be based on the intension of the concepts/terms and not on their extension.

   E. g. if we define "human" as "driver" i.e. by his incidental property to drive a car, then all people who do not drive are not human!

2. In defining the target terms, both the semantic and syntactic ambiguity and vagueness should be avoided. In other words, an expert should be able to decide if some item of his discourse is an instance of the term or not. The items for which such a decision cannot be made should be marginal.

In building and designing thesauri we often encounter two kinds of ambiguity and vagueness:

a) ambiguity related to the substance of a target term (broader categories-top level concepts): the meaning of a target term could be so broad and relative that it could comprise any kind of items without any semantic contiguity or relation.

E.g. what is not a “research object”?

In that case we have to define the broader categories by attributing properties which enable the identification of items *not belonging* to those categories.

b) ambiguity related to the polysemy of a term: a term could have multiple, incoherent meanings related to the semantic field (context) they refer to.

E.g. Mercury could mean: a metal, a planet, a God in mythology

Therefore, before classifying the term we have to define the functional restrictions of the thesaurus (below, step 1), in other words, to clarify the context in which we find the term in order to disambiguate it.

3. The definitions of the target terms should allow us to identify the common meaning and not the boundaries between the source terms.

   E.g. If we define “armed conflicts” as mutually excluding “peaceful conflicts”, we cannot generalize over all the stages in between, and conflicts which evolve into violence.

4. The definitions of the target terms we build in order to subsume the source terms should not be, as much as possible, limited to or dependent on a specific context of use.

   E.g.: X-Ray systems are used by many disciplines, such as medicine, material assaying, art conservation, archaeology. Classifying them as “medical instruments” or “archaeological instruments” would not render anything about their nature. In contrast, they are “instruments for structure analysis of solid things” by substance, rather than by accidental use.

5. A term may be subsumed under multiple broader categories.

   E.g. “carmine” is a “natural dye” and “red colorant”

6. The arrangement of the top level concepts of thesauri and also any expansion of them, either horizontally (through the addition of new top-level concepts) or vertically (through the specialization of the existing ones), must follow the principle of exclusion of contradictions (clash-free expansion of the thesaurus).

Below are some of the examples that show kinds of contradictions we encountered in the environment of controlled vocabularies:

a) terms are defined through self-contradictory properties.

E.g. “Confrontations, conflicts”: “This term comprises complex intentional activities (a combination of activities) that presuppose at least two actors or groups of actors, who understand their interests and demands as competitive and thus aim at their satisfaction through their involvement in situations of controversy”[6] **resulting from natural phenomena.**
! Here, the contradiction lies in the fact that we define the term both as the intentional actions carried out by at least two actors and at the same time as situations resulting from natural phenomena.

b) a broader term A is subsumed under a narrower B, which implies that the narrower term B has less properties than the broader term A. In this case the contradiction lies in the inconsistency with the IsA relationship which is the basis for building hierarchies and requires that subsumed terms (narrower terms) have at least all properties of the subsuming one (the broader term) (see below)

E.g.: broader term: Stelae (it is a concrete piece of stone bearing inscriptions that can be transferred), narrower term: mobile objects (it comprises objects that can be transferred).

[6] Dariah EU, VCC3. To be published.

! Here the contradiction lies in the fact that, the narrower term (mobile objects) does not necessarily inherit *all* the properties of the broader (Stelea).

c) a narrower term is attributed properties of which, at least one excludes the necessary properties attributed to its broader term.

E.g.: broader term: Immobile objects (it comprises objects that cannot be transferred), narrower term: Stelae (it is an object that can be transferred).

! Here the contradiction lies in the fact that "Stelae" seems to posses contradictory properties: it *can* and *cannot* be transferred!

7. The subsumption of narrower under broader terms should be formulated as an inference supporting the inheritance of the properties of potential instances of the broader term to all the instances of the narrower terms (IsA relationship). Otherwise, we fall into the kind of contradiction mentioned above (7b).

E.g. if we define the broader term: "confrontations, conflicts" as: "complex activities (a combination of activities) that presuppose at least two actors or groups of actors, who understand their interests and demands as competitive and thus aim at their satisfaction through their involvement in situations of controversy[7]", then, each of its narrower terms (coups d'etat, legal actions, wars, revolutions, strikes), must inherit the above mentioned properties of the broader term.

Development Steps

Below we describe the four-step recursive process we use for the development of the upper-level concepts of thesauri. After completing a step, we may need to revisit

[7] Dariah EU, VCC3. To be published.

a previous one in order to refine or rework it. This is no problem in itself, as long as the work shows convergence to better stages of knowledge.

Step one: Define the functional restrictions of the source terms.

Starting point: *Source terms*

Source terms may be context oriented and thus present a specific aspect of their meaning. They may also be defined in a vague or subjective way that could lead to ambiguities.

How to handle source terms

1. Define the domain of discourse of the source terms. This way the reference and thus the meaning of the source terms will be clarified, since the meaning often varies depending on the context within which it appears.

2. Define the purpose of building a thesaurus. This will make the function and the potential usage of the thesaurus explicit and thus its applicability in certain scientific fields.

Expected Results

At this point we have successfully achieved to define the functional restrictions that reveal the reference and the applicability of the thesaurus we plan to build!

- ❖ It is important in this step to preserve the possibility of extending the thesaurus in other fields of applicability.

What if: the starting point is an existing vocabulary that we want to enrich or to expand in other fields of application?

Source terms: the existing vocabularies and terminologies.

The domain of discourse is already given but must be reconsidered in combination with the purpose of the revision of the existing vocabularies and terminologies. It is possible that it will remain the same.

The purpose of revising the existing vocabularies and terminologies is to rework them in a way that could be interoperable and can be maintained in a sustainable and scalable way or to integrate the existing vocabularies and terminologies into a coherent overarching thesaurus.

Step two: Define concepts by their intensional properties

Starting point: the reference of the meaning and the possible applicability of the source terms

1. Split the term into as many senses as necessary. The multiple interpretations of a term can not be excluded even if we define the domain of the source terms (step 1).

2. Detect the intensional properties of source terms.

    a. Use the bottom up method and analyze the properties of the source terms.

        - Intensional properties are characteristics which express the nature/substance of a concept and provide an unambiguous recognition of an item as belonging to a category.

        - Intensional properties are the necessary and sufficient conditions for belonging to a category and they cannot be replaced without loss of meaning.

    b. Distinguish between the intensional properties and the incidental or context-dependent behaviors of the source terms.

3. Make the recognition of the intensional properties transparent and base it on information accessible to everyone.
    - When sufficient intensional properties are implicit or not commonly accessible, the definitions are replaced by confining descriptions or refer to commonly known phenomena.
4. Based on the intensional properties, deduce the potential properties of the concepts.
    - Potential properties are consequences of the nature of a thing.
    - They may be confined to a category or not.
    - They may appear at some instances at some time.

What if: the starting point is an existing vocabulary that we want to enrich or to expand in other fields of application?

We follow exactly the same procedure as described in the step 2 in order to define the intentional properties of the concepts.

Expected Results

At this point we have successfully achieved an intersubjecitve and cross-disciplinary approach of the source terms!

Step three: find the target terms

Starting point: an intersubjective and interdisciplinary approach of the source terms.

1. Utilize the intensional properties of the source terms to reveal hierarchical relationships that lead to broader categories.
    - Intensional properties, as the general properties which are attributed to the source terms, lead us to successively

uncover the connections between the source terms and their broader ones.

- Intensional properties lead to building hierarchies which depict the subsumptions of narrower to broader terms.
- A broader category should confine the set of items for which a set of relevant potential properties is applicable.

What if: the starting point is an existing vocabulary that we want to enrich or to expand in other fields of application?

Check if the upper hierarchies of the existing vocabularies are build on the basis of the intensional properties and are consistent with the principles and preconditions mentioned above (*principles and preconditions*).

Build new hierarchies, if necessary, under which we could align the existing hierarchies and terms according to IsA relationship.

Expected results

Building broader categories eventually to context-independent levels, and finally to elementary concepts through which we perceive and conceptualize our reality, and which provide elementary notions of identity based on their substance, such as “physical object”. These are the **facets!**

- ❖ It is important to have in mind that the higher categories can not be justified in a logically exhaustive and strict way. We arrive at them intuitively, using common sense and by reducing complex terms and concepts to more primitive ones.

Step four: finalize the target terms

Starting point: reaching broader categories

1. Check if the hierarchical relationships established so far are consistent and aligned with the IsA relationship. Any narrower term must be a specific case of the broader term and able to inherit all its characteristics.

The hierarchical division of the broader terms must present a necessary connection between the narrower and broader terms.

- ❖ It is important to have in mind that levels of hierarchies are never absolute and complete. Even Facets may have generalizations!

2. Finalize the definitions of the target terms by means of recapitulation of the intensional (and eventually potential) properties and connections between the terms within a hierarchical structure that presupposes concrete functional restrictions already defined (see: *step one*).

What if: the starting point is an existing vocabulary that we want to enrich or to expand in other fields of application?

We follow exactly the same procedure as described in step 4 in order to finalize the target terms.

Expected results

An "open world" classification!

If we follow the methodological guidelines mentioned above we end up with a classification system that does not divide the world in closed spheres of meanings according to specific characteristics, but brings to light hidden connections between the terms and establishes concept relationships!

Good and Bad Terms

Which terms should be preferred as target terms?

1. Concepts that allow concluding potential properties from intensional ones.

2. Concepts that confine many potential properties (“behavior”) that can only apply to a particular intension.

    - ❖ Take into account that each refinement of intension may confine or guarantee another set of potential properties.

3. Concepts that enable an “open world”. Forming broader categories based on the intensional properties, enables potential properties (“behavior”) that does not divide the world into disjoint classes.

    Take into account that:

    - ❖ Through generalizing a concept into a broader category it can be ensured that this concept possesses more general intensional (and potential) properties, **possibly** together with other concepts under that category.

    - ❖ All items (terms, classes) which ***are not*** included in a broader category are not characterized by the intensional properties of that category.

Which terms should be avoided as target terms?

1. Concepts defined by potential relationships since they are, to a great degree, incidental. Not only they do not reveal the essential properties of a term, but also no further independent properties can be derived.

2. Concepts defined by the criteria of particular context of use, context of interest, spatiotemporal contexts are not suitable for indexing.

    - ❖ Take into account that particulars (gazetteers, person lists) are NOT terminologies (but other KOS)

3. Concepts defined by negation (antonymity, complements): In an open world “having not a property” does not imply anything.

4. Concepts that are selected according to the criterion of the affinity of the meaning (content of the terms).

Appendix

Step one: Define the functional restrictions of the source terms -*examples.*

1. Define the domain of discourse of the source terms:

   - E.g.: the humanities. So, speaking of arts the term “reproduction” can be defined as “an imitation or facsimile of a work of art, esp of a picture made by photoengraving or a reproduction portrait” while in biology can be defined as “any of various processes, either sexual or asexual, by which an animal or plant produces one or more individuals similar to itself” (from: The free dictionary, http://www.thefreedictionary.com/)

2. Define the purpose of building a thesaurus:

   - E.g.: the purpose of building a thesaurus is to facilitate a successful search of the existing knowledge.

Step two: Define concepts by their intensional properties- *examples*

1. Split the term into as many senses as necessary. The multiple interpretations of a term can not be excluded even if we define the domain of the source terms (step 1).

   - E.g.: “museum” is an institution but also a building.
   - E.g. “theater” as a building and also as a performance.
   - E.g.“Greece” is referred to the state, but also to the people or the geographic region.

2. Detect the intensional properties of the source terms.

   - E.g.: a *bachelor* is defined as 'unmarried man'. Not being married is an essential property of a bachelor, because one cannot be a bachelor unless he is an unmarried man (necessary condition) and any unmarried man is a bachelor (sufficient condition).

- E.g. Mother is defined “a female who has at least one child”. Being a female is a necessary property of mother but is not sufficient. She must also have a child!

3. Make the recognition of the intensional properties transparent and base it on information accessible to everyone.

   - E.g.: a necessary condition for defining human being could be the DNA. But DNA is not accessible to everyone, so we have to refer to other morphological characteristics, which are accessible to everyone in order to define our term.

4. Based on the intensional properties, deduce the potential properties of the concepts.

   - E.g.: potential properties of the bachelor: no children, is male or female (not a child), live alone etc. ***Not confined to bachelor!***

   - E.g.: potential properties of a person: can drive a car***. Not confined to person!***

   - E.g.: potential property of an amphora: can have painted decoration. ***Not confined to amphora!***

   - E.g. potential property of the mother: could be married. ***Not confined to mother!***

Step three: find the target terms -*examples*

1. Utilize the intensional properties of the source terms to reveal hierarchical relationships that can lead to broader categories.

   - E.g.: defining the intensional properties of the term “bachelor” reveals the broader category under which bachelor can be subsumed: that is the concept “man”, since any bachelor must be a man.

Step four: finalize the target terms-*examples*

1. Check if the hierarchical relationships established so far are consistent and aligned with the IsA relationship. Any narrower term must be a specific case of the broader term and able to inherit all its characteristics.

   E.g.: the broader category under which the term “mobile objects” can be subsumed, as revealed by its intensional properties, is that of “material objects”, since any mobile object must be a material object.

Good and Bad Terms

Which terms should be preferred as target terms?

1. Concepts that allow concluding potential properties from intensional ones.
   - E.g.: from the intensional properties of the term “material object” which is weight and expansion we can conclude potential properties as shape, man-made etc.
   - E.g., only persons can make legal decisions. Therefore, “person” is a good concept.
2. Concepts that confine many potential properties (“behavior”) that can only apply to a particular intension.
   - E.g. Material Object” can have weight, elasticity. A “living individual” consumes energy.
3. Concepts that enable an “open world”. Forming broader categories based on the intensional properties, enables potential properties (“behavior”) that does not divide the world into disjoint classes.
   - E.g. The archaeological term “scraper” can be defined for any blade suitable for a scratching process, but “most lithic analysts maintain that the only true scrapers are defined on the base of use-wear (Wikipedia)”. If we use the latter definition, a search for all scrapers in some databases will miss all possible scrapers. Consequently, a user will not

be able to revise this classification for his own purposes, because items are accessible only by the narrowest definition. This is much worse for research than getting back some “actual non-scrapers”.

Which terms should be avoided as target terms?

1. Concepts defined by potential relationships are, to a great degree, incidental. Not only they do not reveal the essential properties of an item, but also no further independent properties can be derived.
    - E.g. if we define a “human” by “can drive a car”, what are all the persons that cannot drive? If we define an “amphora” by “can have painted decoration”, what are all the amphorae that are not painted?
2. Concepts defined by the criteria of particular context of use, context of interest, spatiotemporal contexts are not suitable for indexing.
    - E.g.: the time is not an internal (substantial) property to define the term “epoch”. It could be the case that during the same time span are manifested two different kinds of “epochs” referring to different things (cultural and technical epochs.
3. Concepts defined by negation (antonymity, complements): In an open world “having not a property” does not imply anything.
    - E.g.: female human = not male human. What are hermaphrodites?
    - Do not complete levels by other activities/other objects (“Shoemaking-other activities”, “elephants-non-elephants”).
4. Concepts that are selected according to the criterion of the affinity of the meaning (content of the terms).
    - E.g.: If we select the term “dance” as target term in order to subsume all the source terms that are relevant to this subject (for example the term

"dancer"[8]), we may fall into inconsistencies because "dance" is an activity while "dancer" is a person. The only relationship between them is external, on the basis of the context and not of the essential characteristics of the concepts.

Bibliography

Baader, Fr., Horrocks, I., & Sattler, U. Description Logics. In S. Staab, & R. Studer (Eds.), *Handbook on Ontologies* (pp. 21-43). Berlin/Heidelberg: Springer.

Beneventano, D., Guarra, Fr., Magnani, St., & Vincini, M. (2004). A web service based framework for the semantics mapping amongst product classification schemes. *Journal of Electronic Commerce Research 5*(2), 114-127.

Doerr, M., & Iorizzo, D. (2008). The dream of a global knowledge network – A new approach. *Journal on Computing and Cultural Heritage*, *1*(1), 1-23.

Gruber, Th. R.(1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human Computer Studies* (43), 907-928.

Kramer, R., Nikolai, R., & Habeck, C. (1997). Thesaurus federations: loosely integrated thesauri for document retrieval in networks based on Internet technologies. *International Journal on Digital Libraries* (1), 122-131.

Millikan, R. G. (2010). On knowing the meaning; with a coda on swampman. *Mind* (119/ 473), 43-81.

Ranganathan, S.R. (2012). *Colon Classification*. N. Delhi: Ess Ess Publications.

Smith, B. (2003). Ontology. In L. Floridi (Ed.), *Blackwell Guide to the Philosophy of Computing and* Information (pp. 155-166), Oxford: Blackwell.

[8] (www.openfolklore.org/et/tree.htm)

Soergel, D. (1998). Thesauri for knowledge-based assistance in searching digital libraries. In: *Proceedings of the 2nd European Conference on Digital Libraries*, Heraklion, Crete.

Sowa, J F. (2000). *Knowledge Representation. Logical, Philosophical and Computational Foundations*, Pacific Grove CA: Brooks Cole Publishing.

Svenonius, E. (2003). Design of controlled vocabularies. In: *Encyclopedia of Library and Information Science* (pp.822-838).N. York: DOI:10.1081/E-ELIS 120009038.

Taylor, A.G. (2006) *Introduction to Cataloguing and Classification*. Westport CT: Libraries Unlimited.