Preprint of a paper to be published in: M. Donnelly and G. Guizzardi (eds.), *Formal Ontology in Information Systems. Proceedings of the Seventh International Conference (FOIS 2012)*, IOS Press, 2012, 133-146.

A method for re-engineering a thesaurus into an ontology

Daniel KLESS[a,1], Ludger JANSEN[bc], Jutta LINDENTHAL[d], Jens WIEBENSOHN[b]
[a] *University of Melbourne*
[b] *University of Rostock*
[c] *RWTH Aachen University*
[d] *Information Consultant, Lübeck*

Abstract. The construction of complex ontologies can be facilitated by adapting existing vocabularies. There is little clarity and in fact little consensus as to what modifications of vocabularies are necessary in order to re-engineer them into ontologies. In this paper we present a method that provides clear steps to follow when re-engineering a thesaurus. The method makes use of top-level ontologies and was derived from the structural differences between thesauri and ontologies as well as from best practices in modeling, some of which have been advocated in the biomedical domain. We illustrate each step of our method with examples from a re-engineering case study about agricultural fertilizers based on the AGROVOC thesaurus. Our method makes clear that re-engineering thesauri requires far more than just a syntactic conversion into a formal language or other easily automatable steps. The method can not only be used for re-engineering thesauri, but does also summarize steps for building ontologies in general, and can hence be adapted for the re-engineering of other types of vocabularies or terminologies.

Keywords. Ontology construction, thesaurus re-engineering, thesaurus reuse

Introduction

The field of applied ontology is longing to build up more and more comprehensive ontologies. As this is a time- and resource-consuming endeavor, it is desirable to shorten the development by adapting existing resources. Such resources could be various kinds of controlled (and sometimes structured) vocabularies that have been developed over decades in various subject areas; often they contain thousands of concepts and are generally considered to be a basis for building complex ontologies. Our interest here is how to re-engineer *thesauri*, a special type of structured vocabulary.

While there is an emerging standard for ontologies in some quarters (like the OBO Foundry [1]), and despite the interest in re-engineering thesauri ontologically, this standard does not carry over to other communities that are concerned with thesaurus re-engineering, as is witnessed by the following four approaches:

(i) Wielinga et al. [2] understand thesaurus re-engineering as treating a thesaurus as a data model: a thesaurus concept such as ‘furniture’ becomes a class in the ontology and properties such as ‘material’, ‘location’, ‘subject’ or ‘title’ are defined for that class so that specific pieces of furniture can be described by precisely these properties. Using

[1] Corresponding Author.

the RDF semantics [3], this approach does not distinguish between instances and classes. Many ontologists, however, see this as an illegitimate confusion between particular and universal entities.

(ii) Soergel et al. [4] put the emphasis of re-engineering on the refinement of the generic relations of a thesaurus, where normally all hierarchy building relations are treated on a par, as well as all associative relationships. Different hierarchical relations have to be distinguished if, e.g., automated reasoning is to be supported (see table 1).

Table 1: Refinement of thesaurus relations according to Soergel et al. [4]

Sub-relations of the hierarchical relation:	Sub-relations of the associative relation:
'Colorado river' instanceOf 'rivers' 'blood' containsSubstance 'blood proteins' 'roots' yieldsPortion 'cuttings' 'Francophone Africa' hasMember 'Benin'	'overgrazing' causes 'desertification' 'plough' instrumentFor 'ploughing'

(iii) In the Semantic Web and Open Linked Data initiatives it is popular to treat the data model of a thesaurus as an "ontology" or "schema", and the thesaurus content as data of that "ontology". While this is mainly a syntactic conversion, the semantic conversion essentially lies for van Assem [5] in differentiating the hierarchical thesaurus relation into two different relationships—a transitive and a non-transitive one. These relationships are then defined as a subtype of the subclass relation in RDFS, although it is recognized that this practice is often incorrect. Van Assem has further suggested using the already existing Simple Knowledge Organization System (SKOS) [6] as an "ontology" (i.e. data model) for thesauri and other vocabularies. The ontology design patterns by Villazón-Terrazas [7] are very similar to the method of van Assem, but solely focus on a syntactic conversion of thesauri and other vocabularies.

(iv) Hepp and de Bruijn [8] define contexts like 'product' or 'service' which can be combined with concepts such as 'TV set' to create categories like 'TV as product' or 'TV as service'. Hepp and de Bruijn see this as sufficient for a script-based creation of "meaningful ontology classes", without really saying what this comes up to.

This overview reveals that there are highly different understandings of what re-engineering thesauri into ontologies means, and what the differences between vocabularies and ontologies are in the first place—and thus also what building ontologies means in general. While (ii), (iii) and (iv) have the advantage of keeping the original thesaurus structure more or less as it is, we do not see how any of these four approaches could create ontologies that can be integrated with other modularly developed ontologies, not to speak of ontologies that provide useful reasoning results alone or in combination with other ontologies—properties that are often considered essential for complex ontologies. Some authors discuss methods or challenges for re-engineering specific thesauri into ontologies that are of interest to us:

(v) Converting the UMLS meta-thesaurus, Hahn [9] and Hahn and Schulz [10] present a method with four steps including a syntactic conversion, removing cycles and inconsistencies, and finally refining and completing formal specifications manually.

(vi) Based on their work on the Finnish YSA and MASA thesaurus, Hyvönen et al. [11] identify missing links in the is-a hierarchy, ambiguity and non-transitivity of the hierarchical relation, the ambiguity of concept meanings as well as the alignment to a top-level ontology as challenges of using thesauri as ontologies.

(vii) Analyzing the NCI Thesaurus, Schulz et al. [12] find that the thesaurus relations cannot be reliably used as ontological axioms.

There have also been approaches to re-engineer vocabularies other than thesauri into ontologies, e.g. the re-engineering of WordNet by Gangemi et al. [13] using what is now known as the OntoClean method [14] or Ontology Design Patterns focusing on the re-engineering of Classification Schemes [15]. These proposals discuss specific aspects or steps of ontological re-engineering, but cannot guide the whole process.

In this paper we suggest a general method for re-engineering a standard-compliant thesaurus into an ontology by making use of top-level ontologies. Doing so, we implicitly detail important differences between thesauri and ontologies and contribute to the understanding of ontology construction. Our focus is a strongly practical one that provides ontology developers with guidelines—not only in terms of modeling principles, but also in how to use the popular description language OWL correctly. We will start with explaining the structure of a thesaurus before we detail the way we derived our method. The method is then presented in various distinguishable steps.

1. The structure of a thesaurus

According to [16], the basic style of thesaurus relationships has been established in 1967 in an appendix 'Rules and conventions' of the Thesaurus of Engineering and Scientific Terms (TEST) [17] and was further developed in international standards [18] [19], recently updated and summarized in the first part of ISO 25964 [20]. As defined in ISO 25964, a thesaurus assembles concepts and terms (in English language often in their plural form), where a thesaurus concept is a "unit of thought", while a term is a "word or phrase used to label a concept". Thesauri establish equivalence relations between terms and hierarchical and associative relationships between concepts:

Equivalence relationships are primarily established between terms that are truly synonymous or are at least treated quasi-synonymous in the thesaurus, e.g. 'diseases' and 'disorders'. Though logically dubious, thesauri can even treat antonyms (opposites) as quasi-synonyms (e.g. 'wetness' and 'dryness'). Quasi-synonymity may also be used to limit the hierarchical depth of a thesaurus, e.g. by treating the terms 'basalt', 'granite' and 'slate' as equivalent to the term 'rock' [20]. One of the equivalent terms is generally chosen as the preferred term to represent the concept for human readers.

The **hierarchical relationship** "should be established between a pair of concepts when the scope of one of them falls completely within the scope of the other. It should be based on degrees or levels of superordination and subordination, where the superordinate concept represents a class or whole, and subordinate concepts refer to its members or parts" [20]. Further, "every subordinate concept should belong to the same inherent category as its superordinate concept, i.e., both the broader and narrower term should represent a thing, or an action, or a property" [20]. However, neither 'class' nor 'category' are explicitly defined in the standard, and the definition of 'scope' as the "semantic boundary of a concept" [20] is not satisfactory.

The standards are more precise in detailing three types of hierarchical relationships that can be distinguished: (a) the generic relationship, (b) the hierarchical whole-part relation, and (c) the instance relationship. The **generic relationship** is defined as "the link between a class or category and its members or species" [20]. Its correctness can be checked by an all-and-some test: e.g., *all* parrots are birds, and *some* birds are parrots. The **hierarchical whole-part relationship** is correctly applied, if the part belongs *uniquely* to the whole (i.e., there can be no whole-part multi-hierarchy). Many uses of "is part-of" in normal language do not hold as general statements. E.g., while

"A wheel is part of a bicycle" is a true normal language sentence, it has no valid equivalence in a thesaurus, since not all types of wheels are part of a bicycle. Such relations can be modeled as associative relations or be addressed by replacing 'wheels' with a more specific concept 'bicycle wheels' [20]. The **instance relationship** is to be applied between a general concept and an instance. Again, the standard does not explicitly define these terms, but examples like 'Alps' or 'Himalayas' as instances of the class 'mountain regions' make clear that instances shall refer to what can be represented by a proper name. The three types of hierarchical relationships can be mixed with each other. They may also be applied in a way that creates poly-hierarchies, i.e. hierarchies where one concept has multiple parents in the hierarchy.

The **associative relationship** is used for "suggesting additional or alternative concepts for use in indexing or retrieval" and it is to be applied between "semantically or conceptually" related concepts that are not hierarchically related [20]. Further, the standard states that whenever one term is used, "the other should always be implied within the common frames of reference shared by the users of the thesaurus. Moreover, one of the terms is often a necessary component in any explanation or definition of the other" and a term should always be "strongly implied by the other" [20].

Other elements that may be relevant for ontological re-engineering are facets and thesaurus arrays, introduced in [19]. **Facets** can occur as the top-level elements in a thesaurus, but may also be used to indicate "breaks" in the thesaurus hierarchy. For example, *'people'* may be used as a facet label under the concept 'agricultural industries' to visually collocate concepts like 'farm managers' or 'shepherds', implying that these concepts are *not* hierarchical subordinates of 'agricultural industries' [20].

Array labels, as we will call them here, indicate the characteristics of division for groups of sibling concepts (called thesaurus arrays). For example, the array label *'by source animal'* groups 'cow milk' or 'sheep milk' as subordinates of 'milk', while the array label *'by fat content'* groups 'full milk' and 'low fat milk' as a separated group of subordinates of 'milk'. The array labels are ontologically interesting since they may give a hint on additional properties specializing subordinate concepts in the is-a hierarchy of an ontology [20].

It should be noted that thesaurus standards have been developed and changed over time, whereas the data structure of actual thesaurus systems is practically inert after they have been implemented. Thus domain-specific thesauri may often not have adopted the past or recent changes in the standards. Moreover, thesaurus designers often neglect existing standards, and it can only be evaluated manually to which degree a specific thesaurus adheres to the rules and guidelines that are provided in the standards [21]; such an evaluation forms step 1 of our re-engineering method.

2. Derivation of the method

The aim of the re-engineering method suggested in this paper is to create ontologies that can be integrated, although they are modularly developed. For this purpose, we will use existing top-level ontologies and well-defined formal relations, strictly separate classes from individuals, and use a formal language for describing the domain.

The developed method is based on insights about the structural differences between thesauri and ontologies. The adopted methods and best practices are also known as the "realist paradigm" described by Smith & Ceusters [22] or Jansen & Schulz [23]; which were used together with the description language OWL. We

consider this approach to be most useful for the working ontologist in a specific domain, as OWL enjoys tool and reasoning support that many other logic-based languages still lack—a factor that is highly relevant in constructing complex ontologies. Our method is equally applicable when using other logic-based languages with potentially greater expressivity, though.

Various steps in our method were refined or emerged from applying the initial set of methodical steps in a re-engineering case study. In this case study we converted an excerpt of the AGROVOC thesaurus concerned with agricultural fertilizers into an ontology. We refrain from providing a detailed account of this case study here—not only because of its complexity, but also due to ongoing work in its refinement of details. We will, however, refer to the case study to illustrate the various steps of our method. Due to its better readability we used OWL Manchester syntax which is also supported by the ontology editor Protégé that we have used in our modeling task.

3. Method description

Our method consists of the following steps:

1. Preparatory refinement and checking of the thesaurus
2. Syntactic conversion
3. Identification of membership conditions (in natural language)
4. Choice and alignment to top-level ontologies and formal relations
5. Formal specification of classes
6. Dissolving poly-hierarchies
7. De-coupling of independent entities
8. Adjustment of spelling, punctuation and other aspects of class and property labels

In the following subsections we will discuss the motivation for each of these steps, an explanation of the activities involved and an example taken from the AGROVOC thesaurus. While the presented sequence of steps is deliberately chosen and advisable to follow, the re-engineering will inevitably be an iterative process that requires the frequent repetition of steps.

3.1. Preparatory refinement and checking of the thesaurus

Re-engineering should begin with checking and refining the thesaurus so that further steps can be performed more easily. We assume that the thesaurus to be re-engineered is in agreement with the ISO thesaurus standards described in section 2, particularly with respect to the distinction between concepts and terms—either by design or by adaptation, which should generally be automatable. As laid out in section 2, hierarchical relationships summarize a variety of ontologically different relationships. As a first step, the generic, hierarchical whole-part and instance relationships of the former thesaurus must explicitly be distinguished. In this course, thesaurus relationships may be detected that are not conformant with the semantics of the relation defined in the thesaurus standards and should not be transferred into the ontology. For latter steps it is also worth introducing array labels where different characteristics of division can be identified.

The thesaurus should also be analyzed for cyclic hierarchical relationships and orphans—unconnected concepts that do not form part of the thesaurus hierarchy. While such cycles and orphans are considered erroneous in thesauri as well, they cannot be

accepted in the ontology at all since former bear a logical contradiction and latter would appear as top-level classes. Orphans should be assigned an appropriate place in the hierarchy or treated as synonyms of existing concepts in the thesaurus. Cycles are best addressed in connection with step 4 of our method.

Example: As many thesauri, AGROVOC does not distinguish between different types of hierarchical relationships. But, as it happens, our analysis revealed that all hierarchical relationships between 'fertilizer' and its subordinated concepts are proper generic relations. Other parts of the AGROVOC thesaurus display the other types of hierarchical relationships like the instance relationship (Colorado River—Rivers) or the part-of relation (Root hairs—Roots). In addition, our analysis revealed several characteristics of divisions relevant for fertilizers (highlighted in italic font) e.g.:

- *by type of dominating plant nutrient* (Nitrogen fertilizers, Potassium fertilizers)
- *by number of plant nutrients* (Compound fertilizers, Nitrogen phosphorus fertilizers)
- *by nutrient release time* (Slow release fertilizers)
- *by organic/inorganic* (Organic fertilizers, Inorganic fertilizers)
- *by aggregate state* (Liquid fertilizers, Liquid gas fertilizers)

3.2. Syntactic conversion

Syntactic conversion aims at representing the thesaurus in a formal language so that it can be further modified in an ontology editor. For this purpose it is necessary to (preliminarily) ignore any semantics of the relations in a formal language and to use that language to store the thesaurus. Since we assumed the thesaurus to be concept-based according to ISO 25964 [20], all thesaurus concepts and facets, if any, will be represented as classes in OWL. The terms of a thesaurus and the labels of the facets become labels of classes. Language tags allow distinguishing the languages of the labels. Subtypes of labels need to be defined, if it is desired to keep the distinction between preferred and non-preferred terms. Definitions, scope notes, and other notes and housekeeping information can be transferred to comments or custom subtypes of such. It might also be desirable to keep array labels as classes with attached labels for ontology maintenance and navigation purposes, although they do not represent ontologically relevant classes. OWL does not provide a modeling primitive that corresponds to array labels.

The **generic relationships**, which often dominate over the other kinds of hierarchical relationships in thesauri, are adopted as is-a relations in the ontology and represented through the subclass axiom. Nevertheless, they are preliminary and likely subject to smaller or more fundamental changes in connection with steps 4-7.

Hierarchical whole-part relations in thesauri *may be* useful in ontologies, if they contribute to the formal specification of classes (see step 5). For this reason they should be preliminarily modeled as simple part-of relations and considered subject to a later re-assessment. The relations are then also subject to potential further refinement depending on the set of formally defined relationships that shall be adopted (see step 4).

The **instance relationships** in thesauri may correspond to relationships between an individual and a class—an assertion that is generally not considered part of the ontology, but rather of a knowledge base. As such it is to be rejected as part of an ontology, acknowledging that knowledge bases can be represented by OWL as well (using the so-called "ABox" of Description Logic).

Associative relationships *may* give hints that there is an ontological relationship between two concepts that contributes to one concept's formal specification as an ontology class. We recommend checking the usefulness of associative relations after step 4 rather than converting them straight into ontological relations here.

It is desirable to carry out the described syntactic conversion automatically, particularly when the goal is to re-engineer an entire thesaurus. In our AGROVOC case study, however, it was most economical to manually transfer the concept 'fertilizers' and its 31 subordinated concepts, all of which became subclasses of 'fertilizer'.

3.3. Identification of membership conditions (in natural language)

In order to prepare the ground for the later formal specification of classes (step 5), we suggest beginning with an informal (natural language) specification of the classes by adding appropriate meta-data. The goal is to identify characteristics that can act as *necessary* and ideally *sufficient* conditions for belonging to the class in question.

The identification of necessary conditions is generally based on natural language definitions. As ISO 25964 neither considers definitions necessary nor offers any rules for definitions, many thesauri do not contain any. Thus, encyclopedia and dictionary definitions have to be referred to. They need to be in line with the meaning of a thesaurus concept that a standard-compliant thesaurus can express through the following means:

- the equivalence relationship between natural language terms—synonyms or quasi-synonyms—representing a concept
- qualifiers to clarify the respective meanings of homonyms— e.g. bank *(financial institution)* or bank *(geography)*
- scope notes—natural language descriptions that restrict or widen the meaning and usage of a thesaurus concept
- the thesaurus hierarchy as a kind of "context" for a thesaurus concept incl. superordinate and subordinate concepts as well as facets or array labels
- associated concepts, some of which indicate what the concept in question is not meant to include—basically a list of "frequently confused concepts"
- explicit definitions.

There exists little practical guidance for how to identify necessary conditions and deciding when a set of sufficient conditions has been identified. In many cases, in particular for most natural kinds, only necessary conditions can be given [24] [25] [26]. In such cases as many as possible necessary conditions should be identified. At some occasions the problem of defining essential properties may have to be solved by making decisions of where to set the limits for borderline cases. At other occasions one may simply accept that no conditions can be defined. In such cases, natural language descriptions should be provided, which are in any case extremely helpful for both ontology maintainer and user. Examples as well as explanations of common misunderstandings should be included in comments, not in definitions.

Example: The AGROVOC thesaurus includes no definitions for its concepts. The 'Fertilizers' concept is located at the following hierarchical position (with "→" representing what is to become the is-a relation):

Fertilizers → Farm inputs → Inputs → Resources

This hierarchy and a dictionary definition of 'resource' [27] suggest that fertilizer is understood as an input to farming, farming being a kind of value production. Our interest was rather in a scientific specification of fertilizer out of social contexts, though. For this purpose the definition in [28] is more fruitful, which allows characterizing a fertilizer as (a) a material (b) participating in (chemical) processes improving the plant nutrition level of soils (c) containing plant nutrients. In the course of further analyses we decided to characterize fertilizer as (a) a material (b) having the disposition to release plant nutrients and (c) containing *sufficient amounts* of plant nutrients.

3.4. Choice and alignment to top-level ontologies and formal relations

Aligning a thesaurus to a top-level ontology and a corresponding set of formal relations includes (a) organizing all the concepts into an is-a hierarchy, (b) asserting the top-level concepts or facets to be equivalent to or subclasses of adequate classes of a top-level ontology and (c) refining the hierarchical whole-part relations according to a set of formally defined relationships (like 'has-abstract-part' or 'grain-of').

While some authors have doubts about the usefulness of top-level ontologies [29], we consider them as a bundle of helpful micro-theories that in their combination take many decisions from the domain modeler, thus easing his burden and securing similar design standards across ontology projects. Moreover, the commitment to an axiomatic top-level ontology such as BFO[2] or DOLCE[3] avoids wrong conclusions and mistakes typical for ad-hoc approaches to modeling ontologies.

Also, a fixed set of formally defined relations (object properties in OWL) ought to be adopted, such as the Relation Ontology[4] or the relations defined in BioTop [30], an upper-domain ontology that can be aligned (through so-called bridges[5]) to both BFO and DOLCE. The adopted relations should have a strong tie with the adopted top-level ontology, because many relations are and should be constrained in their domain and range with reference to a top-level ontology. A useful set of ontological relations will generally comprise spatial, mereological and temporal relations. Most fundamental is the is-a or subclass-of relation, which is a pre-defined part of the Web Ontology Language (OWL).

The generic relationships in a thesaurus are prima facie candidates for becoming is-a relationships in an ontology. Since they may be mixed with hierarchical whole-part relationships in a thesaurus, addressing point (a) may imply re-combining fragments of the thesaurus that are not related by properly applied generic relations. This, in turn, may require introducing new classes to connect these fragments.

The relationships resulting from the alignment to a top-level ontology are still subject of assessment in step 5 of our method. For the assessment of the is-a hierarchy, the OntoClean method [14] may also be helpful. However, it needs further investigation to determine, how much this adds to the alignment to currently existing top-level ontologies described here. For the modeling of other relations (object properties in OWL), several things have to be observed: (1) Any relationship from a class A to another class has the logical force of a necessary membership condition for this class A. (2) Relationships involve implicit or (in OWL) explicit quantification,

[2] http://www.ifomis.org/bfo
[3] http://www.loa.istc.cnr.it/DOLCE.html
[4] http://obofoundry.org/ro
[5] http://www.imbi.uni-freiburg.de/ontology/biotop/

which is relevant for the semantics of relational expressions [12]. Thus, (3) the relationship ‘A isRelatedTo some B’ does not normally imply the inverse relationship ‘B hasRelationFrom some A’. E.g., every bow has as part some bow string, but not every bow string is part of some bow.

Special consideration should be given to poly-hierarchies. As described in [31] and [32], ontologically “correct” poly-hierarchies in the sense that they do not conflict with any of the other rules in our method are rare in practice. Frequently, the existence of poly-hierarchies indicates mistakes in the is-a hierarchy which can cause multiple inheritance problems (also called diamond problem). Such a modeling mistake is illustrated by the two paths for ‘Renewable energy’ in the AGROVOC thesaurus:[6]

Renewable energy → Renewable resources → Natural resources → Resources
Renewable energy → Energy → Physical phenomena → Phenomena

“Renewable energy” may refer to the potential wind or sunshine of an area as a resource, as the first path says. It cannot be energy at the same time, though, because energy is never renewable in a literal sense, as it is, in fact, never lost but only transformed. Thus the poly-hierarchy is easily dissolved by deleting the second path which is incorrect anyway. There are however ontologically correct poly-hierarchies [33]. It is these hierarchies that are addressed in step 6, where also an example is provided.

Example: For modeling agricultural fertilizers we used BioTop [30], an upper-domain ontology that can be aligned (through so-called bridges) to both BFO and DOLCE[7]. The usefulness of BioTop for us lies in its fine-grained distinctions of material entities and its comprehensive set of formally defined relations (object properties).

In our case study we only aligned the thesaurus concept ‘Fertilizers’ and its subordinate concepts to BioTop. Since these AGROVOC concepts are ordered hierarchically by the generic relationship only, we were able to transform these into is-a relations in our fertilizer ontology. In connection with its formal specification (see next step), we defined the class ‘fertilizer’ to be a subclass of the BioTop class ‘compound of collective material entities’. Collective material entities are amounts of molecules. Compounds of collective material entities represent the combination of several “pure” materials. We must declare ‘fertilizer’ to be such compound since there is hardly any pure fertilizer material in real-life environments. Instead, there will always be contained other substances—at least in minimal amounts—and we want to include these under the material we specify here.

3.5. Formal specification of classes

The formal specification of a class through necessary conditions is realized by adding the restrictions identified in step 3 as anonymous superclasses using the subclass axiom. It is then called a *primitive class* in Protégé. The specification of a class through an essential property is realized by adding the necessary and sufficient conditions as anonymous *equivalent classes* using the equivalent class axiom. It is then called a *defined class* in Protégé [34].

The asserted is-a relation is of considerable importance since all formal specifications of superordinate classes are inherited, allowing for most economic

[6] http://aims.fao.org/en/pages/594/sub?mytermcode=25719
[7] http://www.imbi.uni-freiburg.de/ontology/biotop/

specifications of a single class. Instead of defining new classes that are needed in formal specifications, other available ontologies ought to be referenced that have already formally specified these classes. It needs also be assured that no such new classes are created that overlap with what is already closely described by concepts of the imported thesaurus.

Example: The three necessary and sufficient conditions listed in step 3 translate into the following formal specification:

fertilizer SubClassOf 'compound of collective material entities' and
('bearer of' some 'plant nutrient release disposition') and
('has component part' some ('has granular part' some 'plant nutrient molecule'))

As this characterization does not require that a compound material in question contains *sufficient* amounts of plant nutrients, we did not formulate it as a definition of 'fertilizer', but as a necessary condition for being a fertilizer only.

3.6. Dissolving poly-hierarchies

In order to get an ontology that can easily be maintained, poly-hierarchies should be avoided in the asserted ontology. Dissolving poly-hierarchies requires a decision as to which one of two or more hierarchical class paths shall be retained, i.e. which single superclass a "target class" at the bottom end of a poly-hierarchy shall keep. The other class paths are "dissolved" in the sense that (a) the restrictions of the classes along the dissolved class-paths are added to the specification of the target class and (b) any subsumption of the target class under classes of the dissolved class path is removed from the specification of the target class.

Dissolving poly-hierarchies in the *asserted ontology* in such way is one aspect of the "normalization" method recommended by Rector [31]. Notably, the methodical step never results in any loss of semantic information. The poly-hierarchies can later be automatically restored through inference by reasoning algorithms and then become visible in the *inferred ontology* again.

Example: The class 'liquid fertilizer' may initially be subsumed under two superclasses (illustrated in figure 1). It is a defined as follows:

'liquid fertilizer' EquivalentTo (fertilizer and 'portion of heterogenous liquid')

We decided to resolve the poly-hierarchy by making 'liquid fertilizer' primarily belong to the class 'fertilizer'. Thus, we replaced the hierarchical subsumption under 'portion of heterogenous liquid' (indicated through a dotted arrow in figure 1) by adding the restrictions of that class ('bearer of' some ('quality located' some 'liquid value region')) to the specification of 'liquid fertilizer'. Of course, conditions that are already part of the 'liquid fertilizer' specification and its superclasses along the retained class path do not have to be added again to the specification:

'liquid fertilizer' EquivalentTo (fertilizer and
('bearer of' some ('quality located' some 'liquid value region')))

The original hierarchical path will be restored in the inferred class hierarchy.

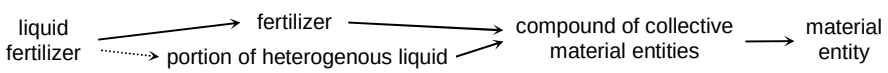


Figure 1. Poly-hierarchy for 'liquid fertilizer' (the dotted arrow indicates the dissolved is-a relation).

3.7. De-coupling of independent entities

De-coupling facilitates the extendibility of ontologies with respect to dealing with non-essential dependent entities—often things like roles, functions or dispositions. Independent entities, the bearers of such dependent entities, are to be modeled independently from their respective combination with one or more dependent entities. The combinations are modeled as defined subclasses of the independent entities.

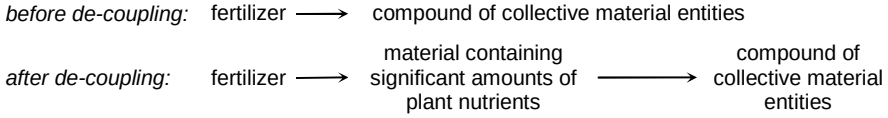
Example: The material through which the class ‘fertilizer’ is described can be separated from its combination with the disposition. Based on the ‘fertilizer’ specification (see step 5), a new class can be introduced:

'material containing significant amounts of plant nutrients' SubClassOf
'compound of collective material entities' and
('has component part' some ('has granular part' some 'plant nutrient molecule'))

Based on this new class, the ‘fertilizer’ specification can be reduced to:

fertilizer EquivalentTo 'material containing significant amounts of plant nutrients'
and ('bearer of' some 'plant nutrient release disposition')

The subsumption hierarchy of ‘fertilizer’ changes as follows:



The design principle of de-coupling independent entities can be used to model additional information about the materials that are the basis for fertilizers. For example, material containing significant amounts of ammonia may not only be used as ammonium fertilizer, but also as cleaner, as agent for food products or as explosive.

3.8. Adjustment of spelling, punctuation and other aspects of class and property labels

The final step of our re-engineering method consists in adjusting the labeling of ontology classes for improving the readability and understandability of the ontology by ontology engineers and users. Currently, there are no universally accepted conventions on how ontology classes should be labeled [35]. Nevertheless, common practices have been summarized [36] and it ought to be checked if similar conventions exist in one’s field. In any case there should be taken care that one style is applied consistently. E.g., it appears to be generally accepted that names for ontology classes should be in their singular form. It should be noted that the labeling described here does not concern the name (URI/IRI) of the classes or properties as specified in [37]. We neither discuss the options for retaining synonym sets from the source thesaurus using the labeling provisions of the respective ontology language.

Example: In all steps above we have already adapted the spelling in the class labels. They now differ from the original spelling of the preferred terms representing the thesaurus concepts (e.g. ‘Fertilizers’ or ‘Calcium fertilizers’).

4. Discussion

Our experience with implementing the re-engineering method is that it requires considerable effort. Steps 2, 6 and 8 may be at least partially automatable while the other steps appear to have no automation potential at the current state of the art. The effort connected with our method is justified where a highly precise standardization of the terminology is required or automated processing of intelligent systems is needed. Search expansion can be improved through ontologies (as well as thesauri), but may not justify the extra effort alone.

Re-engineering thesauri into ontologies requires a variety of ontology-related skills and knowledge just as those discussed for building ontologies [38]. To a large extent, ontology work is a codification of complex domain knowledge. Such knowledge is not at the disposal of non-experts who may need considerable time to acquire the knowledge and may model the domain wrongly. Thus it is desirable to set up a team covering the domain expertise.

Our method assumes an existing thesaurus. Nevertheless, it is only steps 1 and 2 that are specific to thesauri. It may be possible to adapt them to other types of structured vocabularies such as classification schemes (see [39] for an overview of a range of structured vocabularies). The remaining steps can be considered necessary steps for constructing ontologies in general. Steps 3-5 represent the core of ontological modeling. Steps 6 and 7 may be considered optional: Step 6 should have no influence on the hierarchy inferred by reasoning algorithms. And while step 7 explicates ontological principles, it also proliferates the ontology considerably so that it may be perceived more difficult to maintain. Only the specification of classes (step 5) provides clear room for qualitative differences and there may be developed evaluation methods describing those.

We acknowledge that our method has not been tested and refined on the scale of re-engineering a complete thesaurus. This is an extremely time-consuming task of potentially several man-years work given the typical thesaurus size of several thousand or even ten-thousand concepts. The method may have to be adjusted or may be usefully amended based on such large-scale experience.

5. Conclusions

We have presented a method for re-engineering thesauri based on some best practices of ontology construction and thorough consideration of theoretical and practical differences of ontology and thesaurus structures. The method shows that it is not admissible to treat thesauri as ontologies and that there is far more to re-engineering thesauri than a simple syntactic conversion. Obviously, there is a difference between "converting and integrating vocabularies on the Semantic Web" [5] and creating ontologies that can be modularly developed and integrated. Unfortunately, this is often ignored, for example in the Ontology Alignment Evaluation Initiative (OAEI)[8] where thesauri and other vocabularies have been used as test cases in recent years.

Our method creates higher sensitivity towards distinguishing ontologies from vocabularies, sometimes referred to as terminologies. It also contributes to a better understanding of what the construction of complex ontologies means. Though not yet

[8] http://oaei.ontologymatching.org/

an ontology itself, a vocabulary—whether structured or not—is often a good starting point for constructing an ontology. In the re-engineering process, it is the use of logic-oriented modeling with necessary and sufficient conditions that enforces a higher level of precision than is the case with thesauri. Ontological modeling challenges and provokes rethinking of contradictions that stay undiscovered in vocabulary and terminology work. The instruments for specifications are too weak in these contexts.

In future work we plan a more detailed analysis and comparison of the practical differences between thesauri and ontologies. In particular, we want to detail which of those two information artifacts is appropriate for which purpose, in order to know for which aim the effort of re-engineering a thesaurus into an ontology is justified.

Acknowledgements

The research of D.K. has been enabled through the ORES travel and research grant provided by University of Melbourne, with special thanks to Edmund Kazmierczak and Simon Milton for their support in setting up the research visit. The work of L.J. has been supported by the DFG under the auspices of the GoodOD project. We are indebted to Niels Grewe, Johannes Röhl, Simon Milton and Edmund Kazmierczak and the anonymous referees for helpful comments on previous versions of this paper.

References

[1] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, and S. Lewis, "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration," *Nat Biotechnol*, vol. 25, no. 11, p. 1251, Nov. 2007.

[2] B. J. Wielinga, A. T. Schreiber, J. Wielemaker, and J. A. C. Sandberg, "From thesaurus to ontology," in *Proceedings of the 1st international conference on Knowledge capture*, Victoria, British Columbia, Canada, 2001, pp. 194–201.

[3] J. Z. Pan, "Resource Description Framework," in *Handbook on Ontologies*, 2nd ed., 2009, pp. 71–90.

[4] D. Soergel, B. Lauser, A. Liang, F. Fisseha, J. Keizer, and S. Katz, "Reengineering Thesauri for New Applications: the AGROVOC Example," *JoDI*, vol. 4, no. 4, 2004.

[5] M. van Assem, "Converting and Integrating Vocabularies for the Semantic Web," Vrije Universiteit, Amsterdam, the Netherlands, 2010.

[6] SKOS-Reference, *SKOS Simple Knowledge Organization System. Reference*. World Wide Web Consortium (W3C), 2009.

[7] B. M. Villazón-Terrazas and A. Gómez Pérez, "A Method for Reusing and Re-engineering Non-ontological Resources for Building Ontologies," PhD thesis, Universidad Politécnica de Madrid, 2011.

[8] M. Hepp and J. de Bruijn, "GenTax: A generic methodology for deriving OWL and RDF-S ontologies from hierarchical classifications, thesauri, and inconsistent taxonomies," in *The Semantic Web: Research and Applications*, Innsbruck, Austria, 2007, pp. 129–144.

[9] U. Hahn, "Turning Informal Thesauri Into Formal Ontologies: A Feasibility Study on Biomedical Knowledge re-Use," in *Comparative and Functional Genomics*, 2003, vol. 4, pp. 94–97.

[10] U. Hahn and S. Schulz, "Ontology engineering by thesaurus re-engineering," in *Information Modelling and Knowledge Bases Xiii*, H. Kangassalo, Ed. IOS Press, 2002.

[11] E. Hyvönen, K. Viljanen, J. Tuominen, and K. Seppälä, "Building a national semantic web ontology and ontology service infrastructure—the FinnONTO approach," in *Proceedings of the 5th European semantic web conference ESWC 2008, Tenerife, Spain, June 1-5, 2008*, Berlin, Heidelberg, 2008, pp. 95–109.

[12] S. Schulz, D. Schober, I. Tudose, and H. Stenzhorn, "The Pitfalls of Thesaurus Ontologization–the Case of the NCI Thesaurus," in *AMIA Annual Symposium Proceedings*, 2010, vol. 2010, p. 727.

[13] A. Gangemi, N. Guarino, and A. Oltramari, “Conceptual analysis of lexical taxonomies: the case of WordNet top-level,” in *Proceedings of the international conference on Formal Ontology in Information Systems - Volume 2001*, Ogunquit, Maine, USA, 2001, pp. 285–296.
[14] N. Guarino and C. A. Welty, “An Overview of OntoClean,” in *Handbook on Ontologies*, 2nd ed., 2009, pp. 201–220.
[15] ODP, “Proposed Reengineering OP,” *Ontology Design Patterns*. [Online]. Available: http://ontologydesignpatterns.org/wiki/Category:ProposedReengineeringOP. [Accessed: 30-Mar-2012].
[16] J. Aitchison and S. D. Clarke, “The Thesaurus: A Historical Viewpoint, with a Look to the Future,” in *The Thesaurus: Review, Renaissance and Revision*, Haworth Press, 2004, pp. 5–21.
[17] Engineers Joint Council and US department of Defense, Eds., *Thesaurus of engineering and scientific terms (TEST)*. New York: American Association of Engineering Societies, 1967.
[18] ISO 2788:1974, “Documentation - Guidelines for the establishment and development of monolingual thesauri,” International Organization for Standardization, International Standard ISO 2788, 1974.
[19] ISO 2788:1986, “Documentation - Guidelines for the establishment and development of monolingual thesauri,” International Organization for Standardization, International Standard ISO 2788, Nov. 1986.
[20] ISO 25964-1:2011, “Information and documentation -- Thesauri and interoperability with other vocabularies -- Part 1: Thesauri for information retrieval,” International Organization for Standardization, International Standard ISO 25964-1, Aug. 2011.
[21] D. Kless and S. Milton, “Towards Quality Measures for Evaluating Thesauri,” in *Metadata and Semantic Research*, Alcalá de Henares, Spain, 2010, vol. 108, pp. 312–319.
[22] B. Smith and W. Ceusters, “Ontological realism: A methodology for coordinated evolution of scientific ontologies,” *Applied Ontology*, vol. 5, no. 3–4, pp. 139–188, Nov. 2010.
[23] L. Jansen and S. Schulz, “The Ten Commandments of Ontological Engineering,” in *Proceedings of the 3rd Workshop of Ontologies in Biomedicine and Life Sciences (OBML), Berlin, 6.-7.10.2011*, 2011.
[24] S. A. Kripke, *Naming and necessity*. Oxford: Blackwell, 1980.
[25] H. Putnam, “It ain’t necessarily so,” *The Journal of Philosophy*, vol. 59, no. 22, pp. 658–671, 1962.
[26] J. I. Saeed, *Semantics*, 3rd ed. Wiley-Blackwell, 2009.
[27] “Resources,” *International Encyclopedia of the Social Sciences*. Encyclopedia.com, 2008.
[28] V. Gowariker, V. N. Krishnamurthy, S. Gowariker, M. Dhanorkar, and K. Paranjape, *The Fertilizer Encyclopedia*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2008.
[29] D. B. Lenat, “Applied ontology issues,” *Appl. Ontol.*, vol. 1, pp. 9–12, Jan. 2005.
[30] E. Beisswanger, S. Schulz, H. Stenzhorn, and U. Hahn, “BioTop: An upper domain ontology for the life sciencesA description of its current structure, contents and interfaces to OBO ontologies,” *Applied Ontology*, vol. 3, no. 4, pp. 205–212, 2008.
[31] A. L. Rector, “Modularisation of domain ontologies implemented in description logics and related formalisms including OWL,” in *Proceedings of the 2nd international conference on Knowledge capture*, 2003, pp. 121–128.
[32] G. H. Merrill, “Ontological realism: Methodology or misdirection?,” *Applied Ontology*, vol. 5, no. 2, pp. 79–108, Jun. 2010.
[33] I. Johansson, “Four Kinds of ‘Is_A’ Relation,” in *Applied Ontology. An Introduction*, K. Munn and B. Smith, Eds. ontos verlag, 2009, pp. 235–254.
[34] M. Horridge, S. Jupp, G. Moulton, A. Rector, R. Stevens, and C. Wroe, “A Practical Guide To Building OWL Ontologies Using Protégé 4 and CO-ODE Tools Edition1. 2,” *The University of Manchester*, 2009.
[35] V. Svátek, O. Šváb-Zamazal, and V. Presutti, “Ontology naming pattern sauce for (human and computer) gourmets,” in *Workshop on Ontology Patterns at ISWC*, 2009, vol. 9.
[36] D. Schober, B. Smith, S. Lewis, W. Kusnierczyk, J. Lomax, C. Mungall, C. Taylor, P. Rocca-Serra, and S. A. Sansone, “Survey-based naming conventions for use in OBO Foundry ontology development,” *BMC bioinformatics*, vol. 10, no. 1, p. 125, 2009.
[37] RFC 3986, “Internationalized Resource Identifiers (IRIs),” Request for Comments, Jan. 2005.
[38] F. Neuhaus, E. Florescu, A. Galton, M. Gruninger, N. Guarino, L. Obrst, A. Sanchez, A. Vizedom, P. Yim, and B. Smith, “Creating the ontologists of the future,” *Applied Ontology*, vol. 6, no. 1, pp. 91–98, Jan. 2011.
[39] BS 8723-3:2007, “Structured vocabularies for information retrieval. Guide. Vocabularies other than thesauri,” British Standards Institution (BSI), Committee IDT/2/2, 2007.