Tracking Name Patterns in OWL Ontologies

Vojtěch Svátek, Ondřej Šváb

University of Economics, Prague, Dep. Information and Knowledge Engineering,
Winston Churchill Sq. 4, 130 67 Praha 3, Prague, Czech Republic
svatek@vse.cz, svabo@vse.cz

Abstract. Analysis of concept naming in OWL ontologies with set-theoretic semantics could serve as partial means for understanding their conceptual structure, detecting modelling errors and assessing their quality. We carried out experiments on three existing ontologies from public repositories, concerning the consistency of very simple name patterns—subclass name being a certain kind of parent class name extension, while considering thesaurus relationships. Several probable taxonomic errors were identified in this way.

1 Introduction

Concept names in semantic web (OWL) ontologies with set-theoretic semantics are sometimes viewed as secondary information. Indeed, for logic-based reasoners, which are assumed to be the main customers exploiting these ontologies, anyhow cryptic URLs can serve well. Experience however shows that even in ontologies primarily intended for machine consumption, the naming policy is almost never completely arbitrary. It is important for ontology developers (and maintainers, adoptors etc.) to be able to see the semantic structure of a large part of the ontology at once, and ontology editors normally use base concept names (local URLs) and not additional linguistic labels within their taxonomy view. At the same time, while inspecting possibly complex OWL axioms, self-explaining concept names (even independent of their context in the taxonomy) are extremely helpful.

This leads us to the hypothesis that concept naming in OWL ontologies can (at least in some cases) be a useful means for analysing their conceptual structure, detecting modelling errors and assessing their quality. Obviously, a 'true' evaluation of concept naming in specialised domain ontologies requires deep knowledge of the domain. We however assume that even in specialised ontologies, the 'seed' terms often belong to generic vocabulary and the domain specialisation is rather achieved via adding syntactic attributes (such as adjectives or nouns in apposition), leading to multi-word terms. The class-subclass pairs would then often be characterised by the presence of a common token (or sequence of tokens) on some particular position; we can see this as a simple (atomic) *name pattern*. Although the proportion of instances of such a pattern only represent a fraction of all subclass relationships[1], in large- and medium-sized ontologies this may suffice for partial assessment of the consistency of naming, as part of ontology quality evaluation.

[1] Based on our preliminary analysis, we estimate this fraction to float around 50%, depending on domain specificity and other factors.

Atomic name patterns can then be weaved into more complex pattern structures with their own semantics. The deeper understanding of the structure of an ontology thus acquired can help in e.g. mapping it properly to other ontologies.

The paper is structured as follows. Section 2 sets up the token-level background for our name patterns. Section 3 explaines the name patterns themselves. Section 4 describes the initial experiments on three ontologies and attempts to interpret their results. Section 5 surveys some related work. Finally, section 6 summarises the paper and outlines some future work.

2 Matching Tokens in OWL Concept Names

All name patterns we consider in this work are built upon the *sub-string relationships* between pairs of concept names, at token level. The token-level relationship can in general have the nature of prefix, postfix or infix, possibly adjusted with some connective. For example, the name 'WrittenDocument' can be extended via prefix to 'HandWrittenDocument' or via infix to 'WrittenLegalDocument'. A postfix extension could be e.g. 'WrittenDocumentTemplate', which, however, unlike the previous ones, would not be adequate for a subclass of 'WrittenDocument', as the main term (distinguished by its placement as end token) has changed. An adequate postfix extension for a subclass would in turn be e.g.'WrittenDocumentWithComments'; here however the postfix has the form of prepositional construction appended to the main term (thus preserved).

Tokenisation is, for 'technical' items such as OWL concept names, usually assumed to rely on the presence of one of a few delimiters, in particular: underscore (`Concept_name`), hyphen (`Concept-name`) and change of lowercase letter to uppercase (`ConceptName`), which is most parsimonious and therefore most frequent. Although the semantics of these delimiters could in principle differ (especially the hyphen is likely to be used for more specific purposes than the remaining two, on some occasions), we will treat them as equivalent for the sake of simplicity. We will also ignore sub-string relationship without explicit token boundary (i.e. between two single-word expressions), assuming that they often deviate from proper subclass relationship (as in 'fly' vs. 'butterfly', or even worse e.g 'stake' vs. 'mistake').

The mentioned token-level structures then have to be tracked over the *ontology structure* (for simplicity let us only consider taxonomic paths). This could lead to an inventory of *naming patterns*, some of which we considered in our start-up analysis presented below. The most obvious naming pattern is of course the one already mentioned: a subclass name being token-level extension of its parent class. Such patterns can already be assigned some status wrt. ontology content evaluation and possible refactoring. Although the 'token analysis' approach used is admittedly quite naive from the NLP point of view, we believe that, due to the restricted nature of concept names in ontologies, we would not need much more for covering the majority of multi-word names in real-world ontologies.

3 Some Ammunition for Pattern-Based Evaluation

Let us now outline a few, still rather vague, initial hypotheses concerning the interpretation of name patterns.

The first one, concerning subclassing, is central in our initial investigation:

Hypothesis 1 *If the main term in the name of a class and the main term in the* multi-word *name of its immediate subclass do not* correspond[2] *then it is likely that there is a conceptual incoherence.*

The hypothesis anticipates that ontology designers should not often, while subclassing, substantially change the meaning of the main term in the name, as the main term is likely to denote the conceptual type of the underlying real-world entity, and they are obliged to keep the set-theoretic consistency (all instances of the subclass also have to be instances of the parent class). They may however subclass a multi-word name with a rather specific single-word name.

The second hypothesis is closely related:

Hypothesis 2 *If the two main terms from Hypothesis 1 only correspond via some long-range terminological link then it is likely that there is a shift to a more specific domain with its own terminology.*

This hypothesis might help suggesting points for breaking large monolithic ontologies into more and less specific parts.

We also formulated two hypotheses that involve more extensive graph structures of the taxonomy.

Hypothesis 3 *Concept with the same main term in their names should not occur in separate taxonomy paths.*

In other words, if there are several partial taxonomies with the same main term, they are candidates for merger.

Hypothesis 4 *If two taxonomy paths exist such that one contains a class X and its subclass Y, and the other contains a class Z and its subclass W, such that the name of X is token-level extension of the name of Z,* with different main term*, and the name of Y is token-level extension of the name of W,* with different main term*, then both paths should be linked with some property and the name pattern should probably apply for the descendants of Y and W as well.*

This amounts to identification of 'parallel' taxonomies of related (but conceptually different) entities, which may also be quite important e.g. in ontology refactoring as well as mapping.

In the experiments below we only systematically compare Hypothesis 1 to our findings. We however occasionally mention the other three hypotheses where relevant.

[2] The specification of 'correspondance' is discussed in section 4.1.

4 Experiments

In the initial, manual[3], phase of our experiments, we restricted the analysis to 3 small- to medium-sized ontologies we picked from public repositories. Their choice was more-or-less 'random', we however avoided ontologies that appear as mere (converted) ad hoc taxonomies without the assumption of set-theoretic semantics, as well as 'toy' models designed for demonstrating DL reasoning (such as 'pizzas' or 'mad cows'), which are actually quite common in such repositories, cf. [9].

4.1 Settings

In designing the experiments, there were numerous choices, especially concerning:

1. What *patterns* to follow
2. Whether to only consider the own structure of the ontology or also that of *imported* ontologies such as upper-level ones (namely, SUMO, in two out of the three cases)
3. Whether to require for fulfilling the patterns that the main term should be identical in the parent class and subclass, or also allow *hyponymy/synonymy*.

For the first issue we eventually decided to only consider two concrete patterns. The one is the presence of a common *end token*; note that this covers all cases of prefix and infix extension. The second (which proved much more rare) is the postfix extension starting with the 'of' preposition.

For the second issue we decided to restrict the analysis to the *current ontology* only (i.e. both members of the evaluated concept pairs had to be from the current ontology), but including concepts from imported ontologies that belong to the same domain (or mean only very slight domain generalisation). The rationale is that we did not intend to evaluate the way the concepts from the current ontology are grafted on the upper-level ontology, but only the design of the current ontology proper.

For the third issue, we decided to use *WordNet*[4], with the assumption that a general thesaurus is likely to contain the main terms of multi-word domain terms. However, we separately counted and listed the cases where the pattern compliance was established via WordNet only. We did not use WordNet for *single-token* subclass terms[5]; we rather excluded them from the analysis.

The results of the analysis amount to the simple statistics of:

1. Class-subclass pairs where (one of the two considered) name patterns hold directly.
2. Class-subclass pairs where a name pattern holds via WordNet only.
3. Class-subclass pairs where name patterns don't hold even via WordNet, but we eventhough assessed the subclass relationship as correct.
4. Class-subclass pairs where name patterns don't hold even via WordNet, and we assessed the subclass relationship as incorrect (at least at the level of class names).

[3] For examining the ontologies, we simply unfolded their taxonomies in Protégé.

[4] `http://wordnet.princeton.edu/`

[5] Our main focus are specialised domain ontologies, whose single-token terms are likely to either miss in standard lexical databases or exhibit a meaning shift there.

In the tables below, the cases 2, 3 and 4 are explicitly listed and commented. Three symbolic labels were added for better overview. ○ means: correct relationship, contradicts our Hypothesis 1. ● means: incorrect, conforms to our Hypothesis 1. Finally, ⊗ means: main terms correspond via thesaurus, i.e. Hypothesis 1 does not apply[6].

The number of cases 3 ('false positives') and 4 ('true positives') can be viewed as evaluation measures for our envisaged method of conceptual error detection. There could potentially be 'false positives' even among the cases 2 (and theoretically even among the cases 1) due to homonymy of terms; we however did not clearly identify any such case. The *accuracy* of our approach can thus be simply established as the ratio of the number of cases 4 vs. the number of cases 3+4.

4.2 ATO Mission Models Ontology

This, US-based military (ATO probably stands for 'Air Tasking Order') ontology, which we picked from the DAML repository[7], is an ideal example of highly specific ontology rich in multi-token names; there are very few single-token ones, and none of these is involved as subclass in one of the subclass relationships. The ontology contains 86 classes (aside classes inherited from imported ontologies), and there are 116 immediate subclass relationships[8] (including some multiple inheritance). Of them, 95 comply with the name patterns, and 21 don't. Table 1 lists and comments the subclass relationships that break the name pattern. We assume (see the table) that the majority of non-compliance cases (11, i.e. 52%) are modelling errors[9]; some others (5, i.e. 24%) are not strict non-compliance as relationship between the names could be determined using WordNet, and only a few (5, i.e. 24%) seem to be 'false alarms'. In addition, the ontology contains some portions relevant to Hypotheses 3 (e.g. some 'missions' placed beyond the main 'mission' taxonomy and under some other concepts) and 4 (e.g. parallel taxonomies for 'missions' and 'mission plans').

4.3 Government Ontology

This ontology (also from the DAML repository), is relatively smaller and less domain-specific; it contains 53 classes (aside classes inherited from imported ontologies), and there are 27 immediate subclass relationship (including some multiple inheritance). Of the subclass relatioships, 11 comply with the name patterns and 13 don't; finally, 3 involve a single-token subclass, thus being irrelevant for our method. Table 2 lists and comments the subclass relationships that break the name patterns.

4.4 EuroCitizen Ontology

This ontology, picked from the *OntoSelect*[10] repository, contains 71 classes. It has no explicit imports, but largely borrows from SUMO at higher levels of the taxonomy. It

---

[6] But Hypothesis 2 might do if the correspondence is 'long range' only.

[7] `http://www.daml.org/ontologies/`

[8] Here we also considered relationships such that the superclass belonged to the imported but tightly thematically linked ATO ontology.

[9] Or, possibly, artifacts of the DAML→OWL conversion.

[10] `http://olp.dfki.de/ontoselect/`

Superclass	**Subclass/es**	**Comment**
AirspaceControlMeasure	AirCorridor TimingReferencePoint DropZone CompositeAirOperationsRoute	● Subclassing indeed looks misleading. A 'measure' can be *setting up* e.g. a corridor, but not the corridor *itself*.
AirStation	AirTankerCellAirspace	● Rather evokes *part-of* relationship but hard to judge w/o domain expertise.
ATOMission	AircraftRepositioning	○ By the available comment, means AircraftRepositioningMission. However, 'repositioning' looks like acceptable term, though not hyponym of 'mission' in WordNet.
ATOMission	CompositeAirOperations	⊗ 'Mission' is direct hyponym of 'operation' in WordNet. Note however the misuse of plural form.
ATOMissionPlan	IndividualLocationReconnaissanceRequestMission MissileWeaponAttackMission	● The 'Plan' token erroneously missing. The remaining 19 sibling subclasses do have it.
CommandAndControlProcess	AirborneElementsTheaterAirControlSystemMission	● Subclass clearly misplaced, 'mission' concept non contiguous.
CommandAndControlProcess	ForwardAirControl	● Probably means ForwardAirControlProcess.
CommandAndControlProcess	FlightFollowing	⊗ 'Following' could be seen as process (it is hyponym of 'processing' in WordNet). Hypothesis 2 might apply.
ConstraintChecking	RouteValidation	○ Specialisation to subdomain; 'validation' should be closely related to 'checking' but surprisingly is not in WordNet.
ControlAgency	ForwardAirControllerAirborne	○ A tricky case: the end token in subclass is actually an attribute of the true entity ('controller'). Furthermore, although the relationship between 'agency' and 'controller' is not intuitive, it might be OK in the domain context.
ForwardAirControl	AirborneBattleDirection	⊗ 'Direction' is direct subclass of 'control' in WordNet.
GroundTheaterAirControlSystem	ControlAndReportingCenter ControlAndReportingElement	○ Though the relationship between the end tokens is not intuitive, it looks OK in the domain context.
IntelligenceAcquisition	AirborneEarlyWarning	● Rather looks like two subsequent processes: warning is *preceded* by intelligence acquisition. However the end token 'acquisition' bears little meaning by itself.
ModernMilitaryMissile	ArmyTacticalMissileSystem	● A system (i.e. group) of missiles, possibly including a launcher, is probably not a subclass of 'missile'.
PrepositionedMaterielTask	GroundStationTankerMission	⊗ 'Mission' is close hyponym of 'task' in WordNet.
SupportingTask	GroundStationTankerMission	⊗ As above.

Table 1. Name pattern breaks in the ATO Mission Models ontology

Superclass	**Subclass/es**	**Comment**
AreaOfConcern	TransnationalIssue	○ Pattern 2 applies. Interestingly, here the 'semantic' term is rather that *after* 'of': 'issue' is close hyponym of 'concern' in WordNet. Note that Hypothesis 3 would incorrectly suggest to integrate this concept into the taxonomy of geographic areas.
DiplomaticOrganization	ConsulateGeneral	○ A tricky case: the subclass name is a noun phrase obeying French rather than English syntax rules.
GovernmentOrganization	GovernmentCabinet	⊗ 'Cabinet' is hyponym of 'organisation' in WordNet. Hypothesis 2 might apply.
JudicialOrganization	AppealsCourt (+ 3 other court types)	⊗ 'Court' is hyponym of 'organisation' in WordNet. Hypothesis 2 might apply.
LegislativeOrganization	LegislativeChamber	○ Correct. None of the senses of 'chamber' is closely related to 'organisation' in WordNet
OverseasArea	BritishCrownColony UnincorporatedUni- tedStatesTerritory	⊗ Both 'colony' and 'territory' are close hyponyms of 'area' in WordNet.
PoliticalParty	PoliticalCoalition	● Political coalitions often have similar rights as parties but they are not conceptually identical. 'Coalition' is also *not* hyponym nor synonym of 'party' in WordNet.
SuffrageLaw	RestrictedSuffrage	● The (restricted) suffrage by itself is obviously different from the law that imposes it.
SuffrageLaw	VoterAgeRequirement	⊗ With some reservation, voter age requirements can probably be viewed as 'suffrage laws'. This case however reveals the pitfalls of using WordNet, as 'requirement' is indeed *hyponym* of 'law' there. Hypothesis 2 might apply.

Table 2. Name pattern breaks in the Government ontology

Superclass	Subclass/es	Comment
Blood	BloodGroup	● Incorrect. In the veins there are not amounts of a bloodgroup but amounts of blood *having* some group.
CombatSport	MartialArt	○ Correct. Somewhat marginal usage of 'art'.
ContentBearingObject	NaturalLanguage	⊗ 'Object' is again an extremely versatile concept; but 'natural language' is its long-range hyponym in WordNet. Hypothesis 2 might apply.
HumanAttribute	ReligiousBelief	○ Correct. The problem is due to the notion of 'attribute' being extremely versatile.
HumanBloodGroup	RhesusBloodGroupSystem	● Incorrect. The Rhesus system is an individual rather than class; it *defines* blood groups rather than having them as instances.
LandArea	StateOrProvince	⊗ 'Province' is direct hyponym of 'area' in WordNet. However, the term should not be treated as multi-word proper; it is a logical disjunction.
Region	GeographicArea	⊗ 'Area' is direct hyponym of 'region' in WordNet.
TeamSport	IceHockey	⊗ 'Hockey' is hyponym of 'sport' in WordNet. Hypothesis 2 might apply.
WaterSport	InTheWater OnTheWater	● Shortcut that makes the names too context-dependent.

Table 3. Name pattern breaks in the EuroCitizen ontology

is rather heterogeneous (with respect to its relatively tiny size), but contains clusters of related concepts, where name patterns can be identified. The overall quality of the ontology does not seem to be very high, as it contains many clear modelling errors, such as apparent instances formalised as classes. The outcomes of analysis are in Table 3.

4.5 Summary

Table 4 shows the overall figures. The results are obviously most promising for the ATO Mission Models ontology, which is most domain-specific of the three. In general, the proportion of multi-word names seems to decrease with the growing generality of the ontology (EuroCitizen being the most general of the three). The *accuracy* of 'inconsistency alarms', if they were properly implemented, could be acceptable for human inspection and evaluation of the ontology. However, perhaps with the exception of ATO Mission Models, the *coverage* of our simple approach is still too small to guarantee substantial 'cleaning' of taxonomic errors.

	ATO Missions	Government	EuroCitizen
Subclass relationships	116	27	62
with multi-token subclass	116	24	40
Pattern-compliant (identical)	95	11	30
Pattern-compliant (WordNet)	5	8	4
Pattern-non-compliant, incorrect ('true alarm')	11	2	4
Pattern-non-compliant, correct ('false alarm')	5	3	2
Pattern proportion (w/o use of WordNet)	82%	41%	48%
Accuracy of 'alarm'	69%	40%	67%

Table 4. Summary of results

5 Related Work

Our research is to some degree similar to projects aiming at converting shallow models such as thesauri or directory headings to more structured and conceptually clean ontologies [2–6]. The main difference lays in our assumption that the ontologies in question are already intended to bear set-theoretical semantics, and that the 'inconsistencies' in naming patterns are due to either sloppy naming (possibly just reflecting shortcut terminology used by domain practitioners) or more serious modelling errors, rather than being an inherent feature of (shallow) models.

On the other hand, the research in 'true' OWL ontology evaluation and refactoring has typically been focused on their logical aspects [1, 10]. Our research is, in a way, parallel to theirs. We aim at similar long-term goals, such as detecting potential modelling inconsistencies or making implicit structures explicit. We however focus on a different aspect of ontologies: the naming policy. Due to the subtler nature of consistency or implicit structures in these realms (usually requiring some degree of acquaintance with the domain), the conclusions of name pattern analysis have probably to be more cautious than those resulting from logic-based analysis.

6 Conclusions and Future Work

We presented a simple method of tracking name patterns (based on token-level extensions) over OWL ontology taxonomies, which could help detect some errors with respect to their set-theoretic interpretation. Initial experiments on three ontologies from public repositories indicated that the method has some potential, although the performance will probably largely vary from one ontology to another, especially with respect to their domain specificity.

There are various directions in which our current work ought to be extended. First of all, the so far manual process of pattern (non-compliance) detection used in the very first experiments should be replaced by an *automatic* one. We also plan to *reuse* experience from popular NLP-oriented methods of ontology 'reconstruction' from shallow models, such as those described in [3] or [5]. Consequently, we should, analogously to those approaches, adopt at least a simple *formal model*. Furthermore, concept names used as identifiers are obviously not the only lexical items available in ontologies. future (especially, more automated) analysis should pay similar attention to additional, potentially even multi-lingual *lexical labels* (based on `rdf:label`) and *comments*, which may help reveal if the identifier name is just a shortcut of the 'real' underlying concept name. In addition to class names, *property* naming (in connection with their domain and range) should also be followed, e.g. as drafted in [7]. In long term, we perceive as important to combine the analysis of naming patterns with the analysis of *logical patterns*, in the sense of 'guessing' the modeller's original intention that got distorted due to the representational limitations of OWL. Our closely related interest is also the use of discovered patterns for mapping *between* ontologies. We already started to test the behaviour of some well-known (string-based and graph-based) ontology mapping methods with respect to naming patterns present in ontologies, using synthetic ontology-like models [8]. In the future, the analysis of (naming and other) patterns would be used as pre-processing step to mapping.

The research was partially supported by the IGA VSE grants no.12/06 "Integration of approaches to ontological engineering: design patterns, mapping and mining", no.20/07 "Combination and comparison of ontology mapping methods and systems", and by the Knowledge Web Network of Excellence (IST FP6-507482).

References

1. Baumeister J., Seipel D.: Smelly Owls – Design Anomalies in Ontologies. In: Proc. FLAIRS 2005, 215–220.
2. Giunchiglia F., Marchese M., Zaihrayeu I.: Encoding Classifications into Lightweight Ontologies. In: Proc. ESWC 2006.
3. Hepp M., de Bruijn J.: GenTax: A Generic Methodology for Deriving OWL and RDF-S Ontologies from Hierarchical Classifications, Thesauri, and Inconsistent Taxonomies. In: Proc. ESWC 2007.
4. Kavalec M., Svátek V.: Information Extraction and Ontology Learning Guided by Web Directory. In: ECAI Workshop on NLP and ML for ontology engineering. Lyon 2002.
5. Magnini B., Serafini L., Speranza M.: Making Explicit the Hidden Semantics of Hierarchical Classifications. In: Proc. AI*IA 2003.
6. Serafini L., Zanobini S., Sceffer S., Bouquet P.: Matching Hierarchical Classifications with Attributes. In: Proc. ESWC 2006.
7. Svátek, V.: Design Patterns for Semantic Web Ontologies: Motivation and Discussion. In: 7^{th} Conf. on Business Information Systems (BIS-04), Poznan, April 2004.
8. Šváb O., Svátek V.: In Vitro Study of Mapping Method Interactions in a Name Pattern Landscape. Accepted to the Ontology Matching (OM-07) workshop at ISWC 2007, Busan, Korea.
9. Tempich C., Volz R.: Towards a benchmark for Semantic Web reasoners - an analysis of the DAML ontology library. In: EON Workshop at ISWC 2003.
10. Vrandecic D., Sure Y.: How to Design Better Ontology Metrics. In: Proc. ESWC 2007.