Modularisation of Domain Ontologies Implemented in Description Logics and related formalisms including OWL

Alan L Rector
Department of Computer Science
University of Manchester, Manchester M13 9PL, UK
rector@cs.man.ac.uk

ABSTRACT

Modularity is a key requirement for large ontologies in order to achieve re-use, maintainability, and evolution. Mechanisms for 'normalisation' to achieve analogous aims are standard for databases. However, no similar notion of normalisation has yet emerged for ontologies. This paper proposes initial criteria for a two-step normalisation of ontologies implemented using OWL or related DL based formalisms. For the first – "ontological normalisation" – we accept Welty and Guarino's analysis. For the second – "implementation normalisation" – we propose an approach based on decomposing ("untangling") the ontology into independent disjoint skeleton taxonomies restricted to be simple trees, which can then be recombined using definitions and axioms to represent the relationships between them explicitly.

Categories and Subject Descriptors

I.2.4 Knowledge Representation Formalisms and Methods—representation languages.

General Terms

Design

Keywords

Ontologies, OWL, Semantic Web, Description Logics

INTRODUCTION

This paper aims to begin the discussion of methodologies for normalizing ontologies implemented in description logics and related formalisms such as OWL[1] to achieve modularity and easy evolution. The inspiration is taken from normalisation of databases that has long been routine for similar reasons and to avoid update anomalies. Normalised methods for implementing ontologies in OWL and related formalisms are now important because such ontologies are becoming widespread for navigation on the Semantic Web[2], for terminologies and information models in medical records, *e.g.* *Open*GALEN[3] [9], SNOMED-RT/CT[4] [16], and in recent work in bioinformatics *e.g.* [21] and in many other fields. While much other work on ontologies concentrates on general issues of development, *e.g.* [17] or on issues of abstract meaning, *e.g.* [3, 19], this paper concentrates specifically on the engineering issues of robust modular implementation in logic based formalisms such as OWL Furthermore, we concentrate on the domain level ontology rather than the high abstract categories discussed by Guarino & Welty.

The fundamental goal of implementation normalisation is to achieve *explicitness* and *modularity* in the domain ontology in order to support *re-use*, *maintainability* and *evolution.* These goals are only possible if:

- The modules to be re-used can be identified and separated from the whole
- Maintenance can be split amongst authors who can work independently
- Modules can evolve independently and new modules be added with minimal side effects
- The differences between different categories of information are represented explicitly both for human authors' understanding and for formal machine inference.

BASIC CRITERIA FOR NORMALISATION

Rationale

We assume that the basic structure of the ontology to be implemented has already been organised cleanly by a mechanism such as that of Guarino and Welty, and that a suitable set of high level categories are in place. Our goal is to implement the ontology cleanly in as FaCT, OWL, or other logic-based formalism. Such formalisms all share the principle that the hierarchical relation is "is-kind-of"

K-CAP'03, October 23–25, 2003, Sanibel Island, Florida, USA.

[1] http://www.w3.org/TR/webont-req/
[2] http://www.semanticweb.org
[3] http://www.opengalen.org
[4] http://www.snomed.org

and is interpreted as logical subsumption – *i.e.* to say that "B is a kind of A" is to say that "All Bs are As" or in logic notation $\triangleq x.\ Bx \rightarrow Ax$. Therefore, given a list of definitions and axioms, a theorem prover or "reasoner" can infer subsumption and check whether the proposed ontology is self-consistent ("satisfiable").

The list of features supported by various logic based knowledge representation formalisms varies, but for this paper we shall assume that it includes at least:

- **Primitive concepts** *described* by *necessary* conditions
- **Defined concepts** *defined* by *necessary & sufficient* conditions
- **Properties** which *relate* concepts and can themselves be placed in a subsumption hierarchy.
- **Restrictions** *constructed* as quantified role-concept pairs, *e.g. (restriction hasLocation someValuesFrom Leg)* meaning "located in *some* leg".
- **Axioms** which declare concepts either to be disjoint or to imply other concepts.

These mechanisms are sufficient to treat two independent ontologies as modules to be combined by definitions. For example, independent ontologies of dysfunction and structure can be combined in expressions such as "Dysfunction which involves Heart" *(Dysfunction and (restriction involves someValuesFrom Heart))*, "Obstruction which involves Valve of Heart" *(Obstruction and (restriction involves someValuesFrom (Valve and (restriction isPartOf someValuesFrom Heart)))*. Hence complex ontologies can be built up from and decomposed into simpler ontologies. However, this only works if the ontologies are modular. The rich feature sets of modern formalisms such as OWL allow developers a wide range of choices in how to implement any given ontology. However, only a few of those choices lead to the desired modularity and explicitness.

The fundamental observation underlying our proposals for normalisation is based on the truism that logic guarantees that from true premises true conclusions follow. Hence, if the inference algorithms are sound, complete and tractable, then there are only two ways in which a logic based formalism can go wrong: a) the premises can be false; b) the premises can be incomplete – *i.e.* not all information may be represented explicitly.

False premises most commonly result from attempts to work around restrictive formalisms [1]. They are less of a problem with modern formalisms such as OWL using classifiers such as FaCT [5] or Racer [4].

However, incomplete or inexplicit, information remains a problem – most frequently because either a) information is left implicit in the naming conventions and is therefore unavailable to the reasoner, or b) information is represented in ways that do not fully express distinctions critical to the user.

Amongst the distinctions important to users are the boundaries between modules. If each primitive belongs explicitly to one specific module, then the links between modules can be made explicit in definitions and restrictions as in the examples above. However, if primitive concepts are 'shared' between two modules, the boundary through them is implicit—they can neither be separated, since they are primitive, nor confidently allocated to one module or the other. Hence, it matters which concepts are implemented as primitives and which as constructs and restrictions. The key notion in our proposals is that modules be identified with trees of primitives and the boundaries between those trees identified with the definitions and descriptions expressing the relations between those primitives.

Criteria for normalisation of implementations of domain ontologies

We term that part of the ontology consisting only of the primitive concepts the "primitive skeleton".

We term that part of the ontology which consist only of very abstract categories such as "Structure" and "Process" which are effectively independent of any specific domain the "Top level ontology", and those notions such as "Bone", "Gene", and "Tumour" specific to a given domain such as biomedicine the "Domain ontology".

The essence of our proposal for normalisation is that the primitive skeleton of the Domain Ontology should consist of disjoint homogeneous trees. In more detail:

1. The branches of the primitive skeleton of the domain taxonomy should form trees, *i.e.* no domain concept should have more than one primitive parent.
2. Each branch of the primitive skeleton of the domain taxonomy should be homogeneous and logical, *i.e.* the principle of specialisation should be subsumption (as opposed, for example to partonomy) and should be based on the same, or progressively narrower criteria, throughout. For example, even if it were true that all vascular structures were part of the circulatory system, placing the primitive "vascular structure" under the primitive "circulatory system structure" would be inhomogeneous because the differentiating notion in one case is structural and in the other case functional.
3. The primitive skeleton should clearly distinguish:
    a) *"Self-standing" concepts*[5]: most "things" in the physical and conceptual world – *e.g.* "animals", "body parts", "people", "organisations", "ideas", "processes" etc as well as less tangible notions such as "style", "colour", "risk", etc. Primitive self-standing primitives should be *disjoint* but

[5] The phrase "self-standing concepts" is problematic, but has so far produced less controversy than any suggested alternative. In Guarino and welty they correspond to "types", "quasi-types" and certain concepts used to construct representation of "formal and material roles".

open, i.e. the list of primitive children should not be considered exhaustive (should not "cover" the parent), since lists of the things that exist in the world never be guaranteed exhaustive.

b) *"Partitioning" or "Refining"* concepts: value types and values which partition conceptual (qualia- [3]) spaces *e.g.* "small, medium, large", "mild, moderate, severe, etc. For refining concepts: a) there should be a taxonomy of primitive "value types" which may or may not be disjoint; b) the primitive children of each value type should form a disjoint exhaustive partition, *i.e.* the values should "cover" the "value type".

In practice we recommend that the distinction between "self-standing" and "partitioning" concepts be made in the top level ontology. However, in order to avoid commitment to any one top level ontology, we suggest only the weaker requirement for normalisation, *i.e.* that the distinction be made clear by some mechanism.

4. The axioms, range and domain constraints should never imply that any primitive domain concept is subsumed by more than one other primitive domain concept.

Note that requirement 2, that each branch of the skeleton be "homogeneous", does not imply that the same principles of description and specialisation are used at all levels of the ontology taken as a whole. Some branches of skeleton providing detailed descriptors – *e.g.* "forms and routes" of drugs or detailed function of genes – will be used only in specialised modules "deep" the ontology as a whole. Our proposal, however, is that when such a set of new descriptors is encountered, its skeleton should be treated as a separate module in its own branch of the skeleton.

The distinction between "self-standing" and "partitioning" concepts is usually straight forward and closely related to Guarino and Welty's distinction between "sortals" and "nonsortals"[3]. However, the distinction here is made on pragmatic engineering grounds according to two tests: a) Is the list of named things bounded or unbounded? b) Is it reliable to argue that the subconcepts exhaust the superconcept? *i.e.* is it appropriate to argue that "Super & not sub_1 & not sub_2 & not sub_3… not sub_{n-1} implies sub_n"? If the answer to either of these questions is "no", then the concept is treated as "self-standing".

Consequences

The first consequences of criteria 1, 3 and 4 is that all multiple classification is inferred by the reasoner. Ontology authors should never assert multiple classification manually.

The second consequence is that for any two primitive self-standing concepts either one subsumes the other or they are disjoint. From this, it follows that any domain individual is an instance of exactly one most specific self-standing primitive concept.

A third set of consequences of criteria 1 and 3 is that a) declarations of primitives should consist of conjunctions of exactly one primitive (excluding *Thing*[6]) and zero or more restrictions; b) every primitive self-standing concept should be part of a disjoint axiom with its siblings; and c) every primitive value should be part of a disjoint subclass axiom with its siblings so as to cover its value type.

Finally, criteria 4 limits the use of arbitrary disjointness and subclass axioms. Disjointness amongst primitives is permitted, indeed required by criterion 3. However, arbitrary disjointness axioms are almost certain to cause violations of criterion 4)[7] Subclass axioms are allowed to add necessary conditions to defined concepts by causing them to be subsumed by further restrictions, but not to imply subsumption by arbitrary expressions containing other primitives.[8]

Rationale

Minimising implicit differentia

This approach seeks to minimise implicit information. Not everything can be defined in a formal system; some things must be primitive.

In effect, for each primitive, there is a set of implicit notions that differentiate it from each of its primitive parents (the Aristotelian "differentia" if you will). Since these notions are implicit, they are invisible to human developer and mechanical reasoner alike. They are therefore likely to cause confusion to developers and missed or unintended inferences in the reasoner. The essence of the requirement for independent homogeneous taxonomies of primitives is that there be exactly one implicit differentiating notion per primitive concept, thus confining implicit information to its irreducible minimum. All other differentiating notions must be explicit and expressed as "restrictions" on the relations between concepts.

Keeping the skeleton modular

The requirement that all differentiating notions in each part of the primitive skeleton be of the same sort – *e.g.* all structural, all functional etc.– guarantees that all conceptually similar primitive similar notions fall in the

[6] Previously known as "Top" in DAML+OIL and related formalisms.

[7] A stronger criterion concerning disjointness axioms is probably desirable. The only two use cases which we have seen which do not 'tangle' the ontology are a) disjointness between primitive siblings of a common parent; b) disjointness between existential restrictions to represent non-overlap in space, *e.g.* (has_location hasValue Germany) disjoint (has_location hasValue France).

[8] Conveniently, given the other criteria the permitted type of subclass axiom corresponds precisely to those that can be "absorbed" onto primitive conditions and thus have only local, rather than global, impact on classifier performance (see Horrocks, 1998) However, the motivation of the criterion is clarity and modularity. The performance benefits are a welcome added benefit.

Original Hierarchy	Normalised Skeleton Taxonomies
Substance Protein ‘ ProteinHormone’ Insulin ATPase Steroid ‘SteroidHormone’ Cortisol ‘Hormone’ ‘ProteinHormone’ Insulin^ ‘SteroidHormone’ ‘Catalyst’ ‘Enzyme’ ATPase^	… Substance Protein Insulin ATPase Steroid Cortisol … PhysiologicRole HormoneRole CatalystRole **Linking Definitions and Restriction** Hormone Þ Substance & playsRole-someValuesFrom HormoneRole ProteinHormone Þ Protein & playsRole someValuesFrom HormoneRole SteroidHomone Þ Steroid&playsRole someValuesFrom HormoneRole Catalyst Þ Substance & playsRole someValuesFrom CatalystRole Enzyme Þ Protein & playsRole someValuesFrom CatalystRole Insulin → playsRole someValuesFrom HormoneRole Cortiso → playsRole someValuesFrom HormoneRole ATPase → playsRole someValuesFrom CatalystRole

Figure 1: Normalisation of Ontology of Biological Substances and Roles.

same section of the primitive skeleton. Therefore modularisation which follows the primitive skeleton will always include notions that divide along natural conceptual boundaries.

The requirement that the primitive skeleton of the domain concepts form primitive trees is very general and still requires ontology authors to make choices. For example, the notion of the “Liver” might be of a structural unit which serves a variety of functions. It might be classified as an “Abdominal viscera”, “A part of the digestive system”, or a part various biochemical subsystems. One such relationship must be chosen as primary – if we follow the Digital Anatomist Foundational Model of Anatomy[14] or *Open*GALEN [12], we will choose the simple structural/developmental notion that the Liver is an “Organ”. All other classification will be derived from the description of the structure, relationships, and function of that organ. “Liver” will therefore be part of the organ sub-module of the structural anatomy module of the ontology.

Avoiding unintended consequences of changes

New definitions for new concepts can only add new inferences; they cannot remove or invalidate existing inferences. Likewise, adding new primitive concepts in an open disjoint tree can only add information. They may make new definitions and inferences possible, but they cannot invalidate old inferences (*i.e.* cause the ontology to become unsatisfiable). Therefore definitions of new concepts and new disjoint concepts, or even entire disjoint trees, can be added to the skeleton with impunity.

The three operations which can cause unintended consequences are i) adding new restrictions to existing concepts; ii) adding new primitive parents;. iii) adding new unrestricted axioms.

The first – adding new restrictions to existing properties – can be achieved either directly or by adding subclass axioms that cause one class to be subsumed by a conjunction of further restrictions. Adding new restrictions can be partially controlled by domain and range constraints on properties. If the ontology is well modularised, then the properties that apply to concepts in each section of the skeleton are likely to be distinct and therefore unlikely to conflict. The results for existential (*someValuesFrom*) restrictions are almost always easy to predict. They can only lead to unsatisfiability if a functional (single valued) property is inferred to have (*i.e.* “inherits”) two or more disjoint values. Our experience is that in “untangled” ontologies this is rare and that when it does occur it is easily identified and corrected. The results for universal (allValuesFrom) and cardinality restrictions require more care but are at least restricted in scope by modularisation.

However, the second and third – adding new asserted subsumptions between primitives (or expressions involving primitives) or arbitrary axioms asserting subsumption between arbitrary expressions – are completely unconstrained. Hence it is difficult to predict or control what effects follow. Hence the rules for normalisation preclude these constructs even though they are supported by the formalism. Likewise, disjointness axioms can be used as an alternative to negation making the ontology less transparent and harder to understand. Hence there use is confined to the clearly understood case of primitive concepts. In particular the use of constructions such as “A disjoint A” are deprecated as a work around designed to “smuggle” greater expressivity into OWL otherwise restricted formalisms such as OWL-lite.

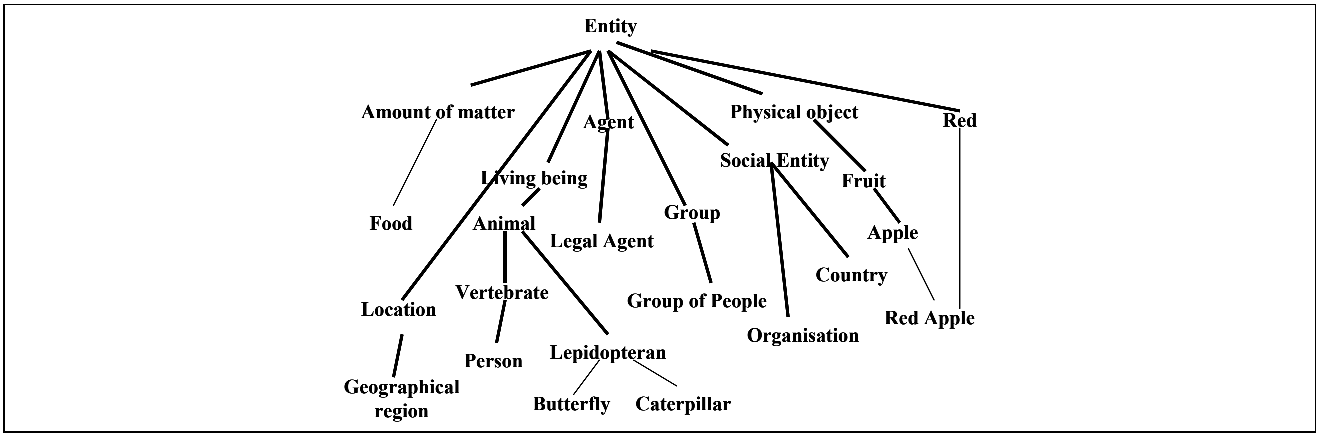
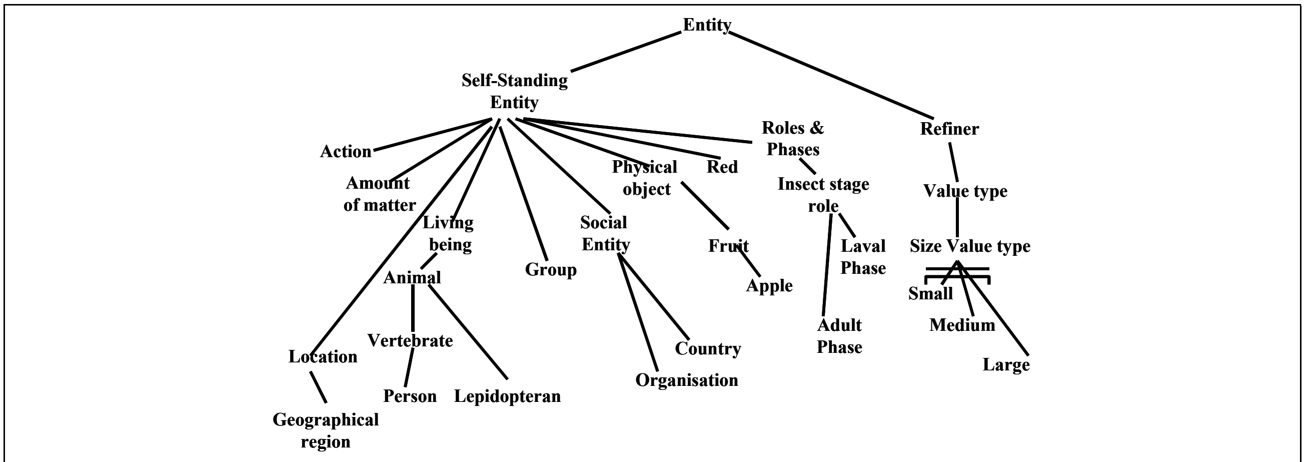


Figure 2a: Example ontology from Guarino & Welty



Agent Þ agentFor someValuesFrom Action

Caterpillar Þ Lepidopteran & (inPhase someValuesFrom LarvalPhase)

RedApple Þ Apple & (hasColour someValuesFrom Red)

BigApple Þ Apple & (hasSize someValuesFrom Big)

Food Þ AmountOfMatter & isEatenBy someValuesFrom Anima

GroupOfPeople Þ Group & (hasMember someValuesFrom Person) & (hasMember allValuesFrom Person)

Figure 2b: Untangled skeleton for example ontology 2a plus definitions linking independent branches

Flavours of is-kind-of

The criteria of normalisation presented here can also be seen as a means to satisfying a common request from knowledge engineers – to be able to have different "flavours" of *is-kind-of.* In effect, we allow exactly one unlabelled flavour of *is-kind-of* link corresponding to the links declared in the primitive skeleton. All others are inferred by the reasoner. In simple cases where they follow from existential restrictions, the restrictions can be thought of as 'labelling' the inferred *is-kind-of* links.

DISCUSSION

Examples & Relation to Other Methods

As a simple example consider hierarchy in Figure 1 for kinds of "Substances". The original hierarchy is tangled with multiple parents for items marked with '^' – "Insulin", and "ATPase". Any extension of the ontology would require maintaining multiple classifications for all enzymes and hormones. Normalisation produces two skeleton taxonomies, one for substances, the other for the physiologic role played by those substances. Either taxonomy can be extended independently as a module – *e.g.* to provide more roles, such as "neurotransmitter role", new kinds of hormone new kinds of protein or steroid, or entire new classes of substances such as "Sugars".

The definitions (indicated by 'Þ') and restrictions (indicated by '→') link the two taxonomies. The resulting hierarchy contains the same subsumptions as the original but is much easier to maintain and extend. (To emphasise the point, the concepts defined in the normalised ontology are shown in single quotes in the original ontology.)

As a further illustration consider the independently developed ontology in figure 2ab adapted from Guarino & Welty (see [3] Figure 6). Figure 2a shows the initial taxonomy after Guarino and Welty's "Ontoclean" process. While ontologically clean, its implementation is significantly tangled. Figure 2b shows the same ontology untangled and normalised.

Each of the changes makes more information explicit. For example, "Food" is classified in the original as part of the backbone simply as a kind of "Amount of matter". In the normalised ontology in Figure 2b, the relation of "Food" to "EatenBy Animal" is made explicit (and the notion of "plant food" therefore explicitly excluded, a decision which might or might not be appropriate to the application but which would likely have been missed in the original. Note also that the nature of the relationship between "red apple" and "red", "big apple" and "big", is now explicit.

The relationship between "lepidopteran", "Butterfly" and "Caterpillar" which causes Guarino and Welty some difficulty as an example of "phased sortals" poses no problem, the relationship of each entity to the generic and to the phase is explicit. Furthermore, general notions such as "group" have been represented explicitly in a re-usable form and ambiguities addressed, *e.g.* Was "group of people" intended as a group *only* of people? *at least* of people? Need a group have any members at all? The normalised representation forces the choice to be explicit rather than leaving it to the individual interpretation of the linguistic label.

Experience

Experience and several experiments support our contention that these techniques are a major assistance in achieving the goals set out in the introduction – *explicitness* and *modularity* in order to support *re-use*, *maintainability* and *evolution.*

This approach to "normalisation", which we also refer to as "untangling", has been used throughout *Open*GALEN and related ontologies over a period of fifteen years [10]. In fact, many of the features of GRAIL, the formalism used in GALEN, were designed around these precepts [7].

Throughout this experience we have found no situation in which the suggested normalisation could not be performed. The requirement to limit the primitive skeleton to simple disjoint trees may seem restrictive, but it does not actually reduce expressiveness. In our experience, violation of this principle almost always indicates that tacit information is concealed which makes later extension and maintenance difficult.

Furthermore, this approach to normalisation or "untangling" has proved easy to explain to new ontology developers and has been one of the key strategies to support loosely coupled development [11]. Interestingly, Gu and her colleagues have independently proposed *post hoc* decomposition into disjoint trees as a means to improve maintainability of large ontologies represented in frame systems with multiple inheritance [2].

We have no comparative data on effort for maintenance, but the combination of normalisation and the user of intermediate representations [9, 11] has allowed us to develop and maintain a large ontology (~30,000 concepts) in a loosely coupled cooperative team consisting at times of up to nine centres in seven countries. The central maintenance and integration effort has been reduced to roughly ten per cent of the total. New modules, for example for methodology and equipment for non-invasive surgery, have been added without incident, almost without comment – *e.g.* it was possible to add the notion of an "endoscopic removal of the gall bladder/ appendix/ ovary/ ulcer/…" in numerous variants to account for different countries' differing practices without any change the modelling of "removal of gall bladder/ appendix/ ovary/ ulcer/…". Furthermore, separate abstractions to provide "views" of the ontology either from the point of view of anatomy or of minimally invasive methodology were quickly and easily constructed and correctly classified and the indexes to the national classifications constructed.

Further evidence for the effectiveness of modularity comes from a study comparing the manually organised UK classification of surgical procedures from Clinical Terms Version 3 (CTv3) with corresponding parts of *Open*GALEN [13]. One source of discrepancies was the inconsistent use in CTv3 of "removal" and "excision" – in some cases removals of a structure were classified kinds of excisions of the same structure; in others the reverse. In *Open*GALEN because ontology is normalised, and "excision" and "removal" are primitives in a module separate from the anatomic structures removed or excised, the same policy is automatically maintained throughout. To take a second example from the same study, another set of discrepancies was traced to minor differences is anatomical boundaries reflecting genuine differences between experts. Each change to the anatomical module in *Open*GALEN could be done in a single place in the anatomy module. Each corresponding change in CTv3 required changes to every surgical procedure concept affected and were widely distributed throughout the surgical procedure model.

Further evidence for the approach comes from the re-use of the *Open*GALEN ontology as the basis for the drug information knowledge base underlying the UK Prodigy project [15]. Perhaps the most dramatic example of the methodology was work on the "simple" problem of forms, routes of administration and preparation of drugs. Although there are only a few hundred concepts, they are densely interconnected and classification had resisted concerted efforts by standards bodies for over two years. Restructuring the classification as a normalised ontology solved the problem in weeks [20].

Issues and problems

The notion of "self-standing"

The notion of "self-standing concept" can be troublesome. In most cases it corresponds to Guarino and Welty's notion of "sortal"; in a few there are questions. For example, consider 'colour'. On the one hand, 'colours' could be considered as partitioning a "qualia space", and the notion of an "identity condition" for colours is problematic. However, in practice, the list of named colours is indefinitely large and constantly growing – witness the efforts of paint companies and interior decorators. To claim a closed list would therefore be inappropriate in most contexts. It is a rare context in which one would be confident in saying "If it is not red or yellow or blue or green… then it must be [say] brown". For most ontologies, we therefore suggest treating colours as "self-standing".

(By contrast, in most contexts we would be happy to accept that "If a measurement is neither low nor normal then it must be elevated". This is true even though we might provide intensifiers such as 'very' or an alternative partition that included "sky high" and "rock bottom". Hence in most ontologies we would recommend that such "modifiers" be treated as "partitioning").

Metaknowledge

A better solution might be argued to be to make the notion of "self-standing" and "partitioning" meta knowledge. These notions are really knowledge about the concepts rather than about all of their instances. Likewise, the notion of whether a concept ought to be part of the primitive skeleton, might be better expressed as metaknowledge. *Open*GALEN and OWL-DL both exclude metaknowledge within the language. Although it is permitted in OWL full, the reasoning support is ill defined. Implementing the distinctions in the ontology itself as suggested here might be considered to be an engineering "kluge" to cope with the limitations of DL classifiers. We would accept this point of view while maintaining the importance of the distinction itself. Hence we advocate that the criterion for normalisation be that there is a means for distinguishing between "self standing" and "partitioning" concepts without specifying the method by which the distinction be made. (A full discussion of the role of metaknowledge in ontologies for the Semantic Web and the OWL family of languages is beyond the scope of this paper.)

Normalisation and Views

The notion of an ontology 'view' is not yet well established. One approach follows database mechanisms and queries [18]. A simpler but useful notion is to provide alternative axes for different uses – structure, function, use, organisational role, etc. If the different modules are clearly separated, then constructing such axes is simply a matter of defining the relevant abstractions, *e.g. BodyPart and hasFunction someValuesFrom F*, for each, or selected, functions *F*. This can also be used in limited circumstances to link to external classifications. In the drug ontology described under "experience", each drug was flagged with its chapter or subchapter in the British National Formulary by simply asserting the restriction *isListedIn someValuesFrom C* for the relevant chapter *C*. Although strictly speaking metaknowledge, this mechanism works pragmatically provided either a) the logical and external classifications are well enough aligned that there are no exceptions, or b) the use case is such that ignoring exceptions can be tolerated. If these conditions are not met, then the indexing mechanisms described below are required.

Indexing & pseudo-default reasoning

This approach has the added benefit that the lattices inferred from normalised, well modularised ontologies provide clean indexes for pointers to other information. These pointers can be used to index information that is not logically implied but nonetheless generally true as in a frame system. Because the ontology is modular, the same type of information is rarely pointed to in more than one branch, hence the set of most specific such pointers usually has only one member. Put another way, "Nixon diamonds" are rare. Hence standard mechanisms for treating defaults and exceptions work relatively well.

The most extensive demonstration of this technique is the PEN&PAD user interface for general practice which has been well validated in repeated human factors studies and was eventually commercialised [6]. PEN&PAD is based indexing fragments of data entry forms using the lattice derived *Open*GALEN The fragments are assembled to construct a data entry form adapted to the particular disease, clinical setting and user preferences. Several hundred thousand highly tailored forms were assembled from a knowledge base of fewer than ten thousand indexed 'facts'. The same techniques appear relevant to assembling information from the semantic web to provide highly tailored presentations and user interfaces.

The other extensive use of this mechanism was in the "code conversion module" of the GALEN terminology server which used such indexing to map from the GALEN ontology to the International Classification of Diseases (ICD9/10) [8]. These mappings were too complex to use the simple view method above. A notable success was that all of the "exclusions" in ICD9/10 – *e.g.* "hypertension excluding in pregnancy" –proved to be cases where there was a mapping to a more specific concept in the GALEN ontology from another ICD9/10 code. They were therefore dealt with simply by the standard mechanism for defaults with exceptions.

Conclusion

The ability of logical reasoners to link independent ontology modules to allow them to be separately maintained, extended, and re-used is one of their most powerful features. However, to achieve this end all information must be explicit and available to both reasoners and authors. The large range of options provided by description logics mean that implementers need guidance on to achieve this end. The approach presented here is based on fifteen year's experience in the development of large (>35,000 concept) biomedical ontologies. The procedures are not an absolute guarantee of a clean, untangled implementation. Not all obscure constructs are completely debarred nor all unintended consequences eliminated, but they are greatly reduced. Others may wish to challenge these criteria or propose further restrictions. However, we believe that if the potential of OWL and related DL based formalisms is to be realised, then such criteria for normalisation need to become well defined and their use routine.

REFERENCES

1. Doyle, J. and Patil, R. Two theses of knowledge representation: Language restrictions, taxonomic classification and the utility of representation services. *Artificial Intelligence*, *48* (1991). 261-297.
2. Gu, H.H., Perl, Y., Geller, J., Halper, M. and Singh, M. A methodology for partitioning a vocabulary hierarchy into trees. *Artificial Intelligence in Medicine*, *15* (1999). 77-98.
3. Guarino, N. and Welty, C., Towards a methodology for ontology-based model engineering. in *ECOOP-2000 Workshop on Model Engineering*, (Cannes, France, 2000).
4. Haarslev, V. and Moeller, R., Expressive ABox reasoning with number restrictions, role hierarchies, and transitively closed roes. in *Pro 7th Int Conf on Knowledge Representation and Reasoning (KR2000)*, (San Francisco, CA, 2000), Morgan Kaufmann, 273-284.
5. Horrocks, I., Using an expressive description logic: FaCT or Fiction. in *Principles of Knowledge Representation and Reasoning: Proceedings of the Sixth International Conference on Knowledge Representation (KR 98)*, (San Francisco, CA, 1998), Morgan Kaufmann, 634-647.
6. Nowlan, W., Rector, A., Kay, S., Horan, B. and Wilson, A., A Patient Care Workstation Based on a User Centred Design and a Formal Theory of Medical Terminology: PEN&PAD and the SMK Formalism. in *Fifteenth Annual Symposium on Computer Applications in Medical Care. SCAMC-91*, (Washington DC, 1991), McGraw-Hill, 855-857.
7. Rector, A., Bechhofer, S., Goble, C., Horrocks, I., Nowlan, W. and Solomon, W. The GRAIL concept modelling language for medical terminology. *Artificial Intelligence in Medicine*, *9* (1997). 139-171.
8. Rector, A., Solomon, W., Nowlan, W. and Rush, T. A Terminology Server for Medical Language and Medical Information Systems. *Methods of Information in Medicine*, *34* (1995). 147-157.
9. Rector, A., Wroe, C., Rogers, J. and Roberts, A., Untangling taxonomies and relationships: Personal and practical problems in loosely coupled development of large ontologies. in *Proceedings of the First International Conference on Knowledge Capture (K-CAP 2001)*, (Victoria, BC, Canada, 2001), ACM, 139-146.
10. Rector, A.L. Clinical Terminology: Why is it so hard? *Methods of Information in Medicine*, *38* (1999). 239-252.
11. Rector, A.L., Zanstra, P.E., Solomon, W.D., Rogers, J.E., Baud, R., Ceusters, W., W Claassen, Kirby, J., Rodrigues, J.-M., Mori, A.R., Haring, E.v.d. and Wagner, J. Reconciling Users' Needs and Formal Requirements: Issues in developing a Re-Usable Ontology for Medicine. *IEEE Tran on Information Technology in BioMedicine*, *2* (1999). 229-242.
12. Rogers, J. and Rector, A., The GALEN ontology. in *Medical Informatics Europe (MIE 96)*, (Copenhagen, 1996), IOS Press, 174-178.
13. Rogers, J.E., Price, C., Rector, A.L., Solomon, W.D. and Smejko, N. Validating clinical terminology structures: Integration and cross-validation of Read Thesaurus and GALEN. *Journal of the American Medical Informatics Association* (1998). 845-849.
14. Rosse, C., Shapiro, I.G. and Brinkley, J.F. The Digital Anatomist foundational model: Principles for defining and structuring its concept domain. *Journal of the American Medical Informatics Association* (1998). 820-824.
15. Solomon, W., Wroe, C., Rogers, J.E. and Rector, A. A reference terminology for drugs. *Journal of the American Medical Informatics Association* (1999). 152-155.
16. Spackman, K.A., Campbell, K.E. and Côté, R.A. SNOMED-RT: A reference Terminology for Health Care. *Journal of the American Medical Informatics Association (JAMIA)* (1997). 640-644.
17. Uschold, M. and Gruninger, M. Ontologies: principles, methods and applications. *Knowledge Engineering Review*, *11* (1996).
18. Volz, r., Oberle, D. and Studer, R., Views for light-weight web ontologies. in *ACM Symp on Applied Computing (SAC-2003)*, (Melbourne, Florida, 2003).
19. Welty, C. and Guarino, N. Supporting ontological analysis of taxonomic relationships. *Data and Knowledge Engineering*, *39* (2001). 51-74.
20. Wroe, C. and Cimino, J., Using openGALEN techniques to develop the HL7 drug formulation vocabulary. in *American Medical Informatics Association Fall Symposium(AMIA-2001)*, (2001), 766-770.
21. Wroe, C., Stevens, R., Goble, C.A. and Ashburner, M., An Evolutionary Methodology To Migrate The Gene Ontology To A Description Logic Environment Using DAML+OIL. in *Pro 8th Pacific Symposium on Biocomputing (PSB)*, (Hawaii, 2003), 624-635.