



Data Registries Technical Note 233480/TN/16

Highways Agency Core Components Analysis

A process for the harmonisation of data concepts

White Paper

Authors: Ian Cornwell and Alastair Dunsmore (Mott MacDonald).

Date: 19 March 2008

1 Purpose and Scope

This paper presents the Highways Agency's Core Components Analysis process as developed by Mott MacDonald and the Highways Agency to encourage harmonisation of data standards, specifications and system interfaces.

The need for harmonisation

In the past decades, many systems have been developed in related and overlapping areas. Due to the range of varying requirements and developer preferences, and a lack of standards, there is much diversity in the methods of data documentation and data representation. The result is the duplication of development and data collection, wasting development costs and losing the efficiency of data interoperability and reuse.

Harmonisation in this context is the process that increases the alignment of data definitions across related systems, leading to benefits in reuse, in interoperability and in development costs. Even if immediate rework of implementations is not an option, the harmonisation process should increase the understanding of the relationship between different data specifications and establish a preferred target for the future.

Assumed context for the technique

The technique described in this paper is used within the Highways Agency's "ITS Metadata Registry" project (www.itsregistry.org.uk), but is described here independently from that context. The technique should work best within an active metadata registry, but the minimum requirements are that all of the data definitions that are in scope are available in a form that allows understanding of their semantics and structure, and that there is a mechanism for publication of results of the harmonisation process.

The technique uses UML for expression. While the technique could be modified for use with other methods of expression, it is particularly suited to diagrammatic presentation.

This document has been prepared for the titled project or named part thereof and should not be relied upon or used for any other project without an independent check being carried out as to its suitability and prior written authority of Mott MacDonald being obtained. Mott MacDonald accepts no responsibility or liability for the consequence of this document being used for a purpose other than the purposes for which it was commissioned. Any person using or relying on the document for such other purpose agrees, and will by such use or reliance be taken to confirm his agreement to indemnify Mott MacDonald for all loss or damage resulting therefrom. Mott MacDonald accepts no responsibility or liability for this document to any party other than the person by whom it was commissioned.

To the extent that this report is based on information supplied by other parties, Mott MacDonald accepts no liability for any loss or damage suffered by the client, whether contractual or tortious, stemming from any conclusions based on data supplied by parties other than Mott MacDonald and used by Mott MacDonald in preparing this report.

This paper assumes that data concepts, however they are represented, may have structure, such that individual property definitions are grouped into aggregate entities representing larger-grained concepts in the subject domain, and these entities may have relationships to one another. This basic idea is common to most description languages and metamodels (for example the Highways Agency project covers XML Schema, CORBA IDL, entity-relationship models, UML models and informal spreadsheets).

2 Challenges in harmonisation

For a single underlying domain concept, there are many types of difference that can arise between the expressions of that concept in two different systems. It is very common for the same conceptual item to occur with different names in different systems or specifications. It is also very common for semantically similar entities to have different boundaries and different structural decompositions in different systems. In general there will be a *set* of entities *partially* corresponding to another *set* of entities.

Even where two systems or specifications apparently have similar scope when viewed at a high level, there may be entities present in one system that are entirely missing in the other. In an example from highway location referencing in the UK, one data model included the following three concepts:



while another system used:



There was an approximate semantic equivalence between “RoadSection” and “Section”, and between “Section_LRP” (which stands for Location Reference Point) and “RoadsidePoint”, but there was no direct equivalent for “Link” – due to differing requirements and business rules about segmentation of the road network. Any harmonisation between the two models has to resolve the issue of how to harmonise the relationship of the “Sections” to the “Points”, which in the second model is direct but in the first model is through the intermediate concept of “Link”. Harmonisation of the “Section” and “Link” entities would also have to resolve the differences in business rules.

Harmonisation has thus to deal with issues at a semantic level, at a structural level, and at a syntactic level. Every part of a data model could potentially vary from system to system even though the same concepts were being described. These parts will include names, attributes, relationships, the boundaries of structures and datatypes. And although the scope of harmonisation is for semantically related concepts, the detailed semantics and business rules may differ and therefore also require resolution.

Harmonisation is easier to achieve if a single organisation owns all of the systems or specifications being harmonised. Harmonisation is particularly difficult in a mature domain where there are already established implementations and standards but no single controlling authority to enforce the use of one particular standard. Nevertheless even within a loosely aligned community a harmonisation process can still be valuable in signalling preferred representations and providing aids to translation or migration.

3 Harmonisation solutions

Harmonisation is made possible and worthwhile because the data models or other data specifications are an expression of similar concepts from the subject domain. Their scope and structure has been influenced by specific business contexts and other contextual factors, but they do reflect concepts from the subject domain. Looking at two sets of data with overlapping semantics but different formats, an analyst is able to attempt to understand similarities because the descriptions show that the semantics overlap – both models are representing similar domain concepts. The analyst is able to identify semantic similarities because he/she has a mental reference model of the subject domain, and can identify that, despite the differences that may exist, the data definitions in each set are a representation of concepts in the subject domain.

In simple cases it is possible to proceed to harmonisation directly without making the background reference model explicit, for example by making immediate changes to the original data specifications. However, this paper asserts that there is value in making the background reference model explicit in the form of *ontology*, i.e. a rigorous conceptual schema representing the subject domain.

The difference in nature between the background ontology and the data specifications undergoing harmonisation must be stressed. Admittedly, the data specifications may already be seen as attempts at rigorous expressions of the subject domain, but they are shaped by their specific business contexts. The ontology used in harmonisation should be more independent of business context. A harmonisation process can then use this reference ontology in the understanding and also the recording of the similarities and differences between different data specifications. The ontology also acts as a reference that should be used as the basis of the target for future developments.

Core components

UN/CEFACT has published the “Core Components Technical Specification” [1], part of the ebXML Framework. The key idea that supports harmonisation is the separation of “core components”, which have no specific business context, from “business information entities”, which apply in specific business contexts. The core components are therefore an example of the background ontology described above.

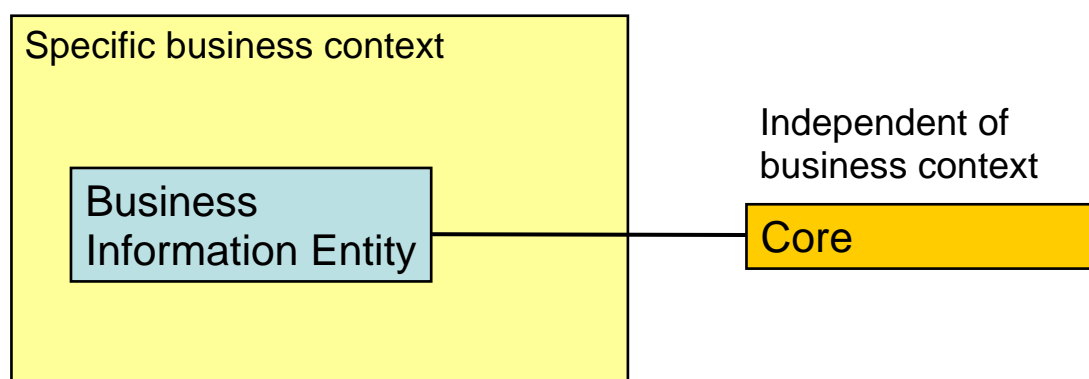


Figure 1 Fundamental idea of core components

Each business information entity is a specific and possibly restricted instantiation of a core component, for a specific business context.

There are different kinds of business information entities and core components: “aggregate” entities are like UML classes, “association” entities are like UML associations, and “basic” entities are like UML attributes. A “property” may be an association or a basic attribute.

The Core Components Technical Specification does not provide fully comprehensive guidance for the harmonisation process. Recognising this, UN/CEFACT TBG 17 developed “Harmonization team submission guidelines and procedures” [2] to provide refined guidance.

Highways Agency Core Components Analysis

The UK Highways Agency has also derived a process of Core Components Analysis to encourage harmonisation of data concepts. This was initially developed independently from the UN/CEFACT TBG 17 guidance, but produced many similarities to that guidance. However, it differs in scope and in detail from the process applied by UN/CEFACT. The UN/CEFACT process aims to ensure global interoperability. The Highways Agency approach is more focussed on incremental improvements to legacy systems and specifications.

The process uses an extended ISO 14817 metadata registry implementation. The submitted data definitions all have their individual business contexts and they can be seen as context-specific instantiations of a single set of underlying concepts, the core components, as shown in the example of **Figure 2**.

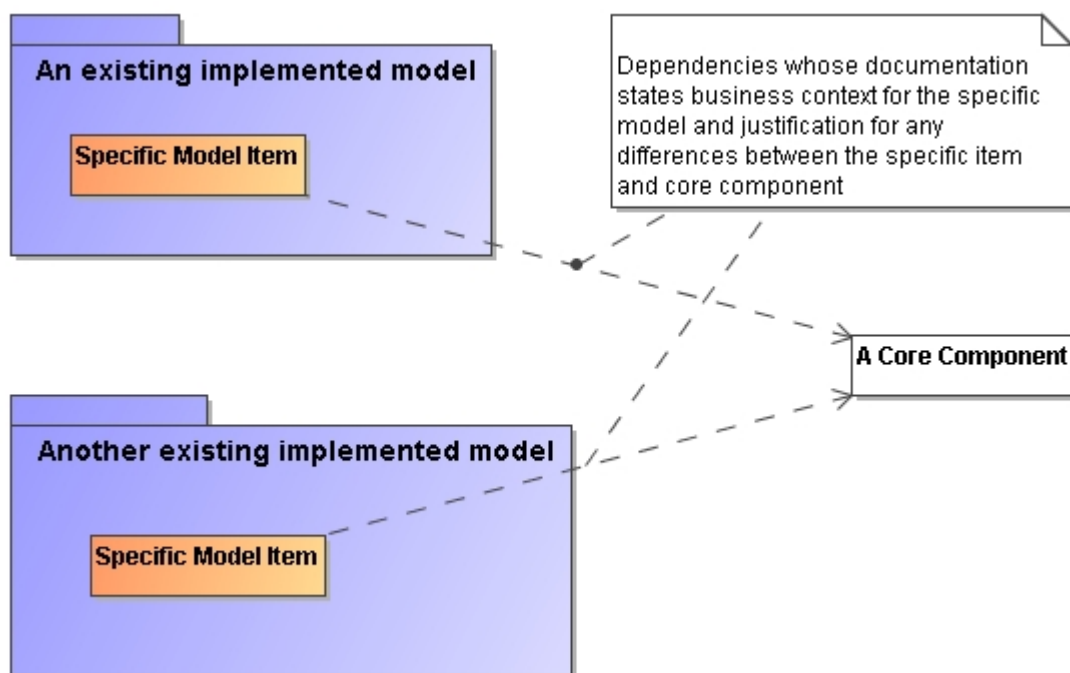


Figure 2 Core component example

The set of core components is derived through analysis of the existing registered models. The process aims to be as objective as possible to avoid the possibility of the core components being yet another competing model. The core components are not introduced unless justified by existing models. The engineer should resist the temptation of “fixing” the registered models unless there are other registered models or established design rules to justify the improvement.¹ Therefore while UN/CEFACT Core

¹ This focus on existing systems, with the avoidance of inventing new design, is reminiscent of practices within “eXtreme programming” [3], and also the technique of “eXtreme Ontology” [4].

Components are supposed to be independent of business context, the core components of the Highways Agency’s metadata registry do have a broad business context that is the superset of all the business contexts of the submitted models. This helps ensure that any analysis effort is constrained to be focussed on Highways Agency business needs.

The outputs of the analysis process are the set of core components plus the mappings from registered models to core components. Each link observed between a registered item and a core component has associated text attempting to explain and justify the differences. If there are legitimate justifications for the shape of the model, those will be declared, but if the model contains poor design or sloppy thinking, these should also be exposed because a good business justification cannot be agreed. This gives a powerful way to provide objective review feedback and improve the quality of submitted models, since there is often no genuine justification and the submitter can understand and agree this. Furthermore the improvements made through this process are in line with requirements brought by other models, and any changes should bring the submission closer to the others.

This process is therefore slightly different from the UN/CEFACT process (depicted in Figure 3), which starts with business requirements and then derives the core components and specific business information entities. The business requirements may or may not have been derived from existing systems.

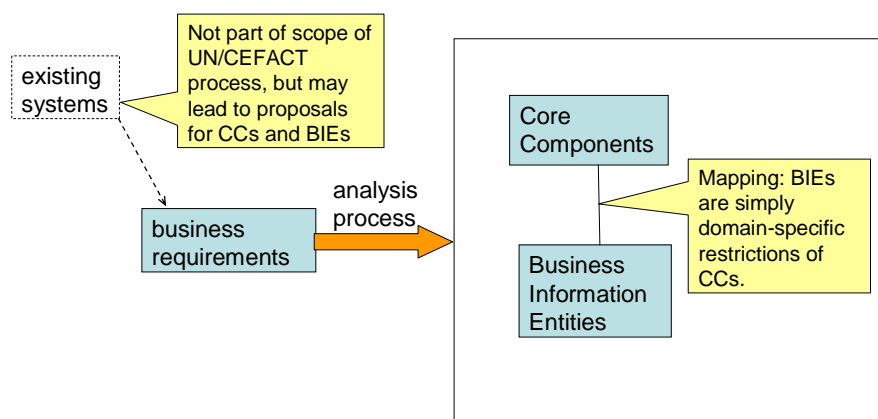


Figure 3 UN/CEFACT core components analysis process

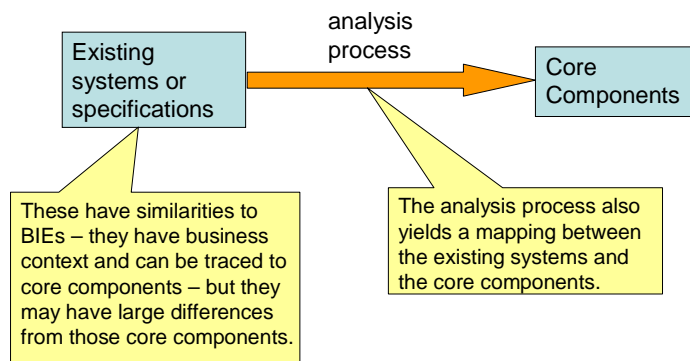


Figure 4 — Highways Agency core components analysis process

In the UN/CEFACT approach, the “Business Information Entities” may be created at the same time as the core components and may differ from the existing legacy systems.

In the Highways Agency approach (depicted in Figure 4) the existing data specifications are expressed alongside the core components and shown to be context-specific instantiations of those components. As the core components are created in response to multiple independently developed systems or specifications, perhaps containing flawed design, the differences between those systems and the core components are greater than those allowed by the UN/CEFACT Core Components metamodel, and so the mapping has to be more flexible. The mappings are expressed as UML dependencies, and each model element from existing specifications may have one or more dependencies to any model elements in the core components model.

In the UN/CEFACT approach, the BIEs provide an explicit target for the future evolution of the existing systems. In the Highways Agency approach, this target is not so explicit, but it is suggested by the combination of the core components and the words of the mappings. In the Highways Agency context, the business case for the effort in creating explicit *ideal* BIEs is not clear as the stakeholders for the submitted specifications may not be receptive. Instead the process looks case by case on the feedback that can be given to stakeholders of existing systems. A compensating advantage of the process is that the mappings from legacy systems to core components are explicit.

The core components analysis is performed for a given subject matter area where harmonisation in that area is of particular business benefit to the Highways Agency. All registered definitions in the subject matter area are considered. The following steps are performed. They are not independent and are likely to be applied in multiple iterations.

1. Build up an outline conceptual schema by including concepts one by one from individual registered data definitions. Show how each submitted entity maps to the core component entities.

2. Fill in the attributes of the core components by considering each submitted attribute. Show how each submitted attribute maps to its corresponding core component. This may lead to an alteration of the entity boundaries from the first step.

The core components provide background ontology. For that purpose alone, the core components could be expressed in an abstract way. However, to encourage harmonisation it is desirable for the constructs used in core components to have a similar level of abstraction to the constructs used in submitted models. The core component attributes therefore use specific concrete datatypes, and the analysis process follows a guiding rule:

If the core components were implemented, they should be able to represent any data currently representable by the submitted models (within the chosen scope areas of the core components) with the added condition that it would be permissible to have empty attributes and associations in the core components implementation.

If the submitted models are similar in approach, then it can be relatively straightforward to create core components. If the subject matter area involves a complex taxonomy, then it may be worthwhile defining how the submitted items map to a conceptual taxonomy before proceeding to the full conceptual schema. If the submitted models take radically different approaches because of justified business requirements, it can be difficult to objectively choose the best way to represent the core components.

The following rules supplement the guiding rule above in the selection of core components. It is interesting that UN/CEFACT TBG 17 also evolved very similar rules independently.

Choose the most general mechanism that preserves all the semantics of the original structure.

While a specific model may allow the representation of a single concept in two ways, for example to allow backwards compatibility, the core components should not give the choice - as long as an implementation of a single mechanism would be capable of conveying the same semantics.

All names must follow a stated detailed style policy. When submitted styles differ, the choices of core component names are therefore made objectively.

Core components analysis can be time consuming, and in particular the registration of the mappings can produce high maintenance if the registered models are not stable. The approach of the Highways Agency project is to proceed with core components analysis only in subject matter areas where further modelling or translation work of relevance to the Highways Agency is imminent or ongoing.

Mappings for submitted models are registered only when those submitted models are relatively stable.

The Highways Agency approach has developed similar ways of portraying the mappings to those developed by TBG 17. Figure 5 shows an example of the mapping of entities and relationships, and Figure 6 shows an example of the mapping of attributes. Occasionally, where there is a complex taxonomy in the subject domain, it can be useful to analyse the taxonomy before proceeding to the full conceptual schema. Figure 7 shows an example of taxonomy mapping. In all these figures the core components are the elements shown as targets of the dependency arrows, while the submitted models are the elements from which the dependency arrows originate.

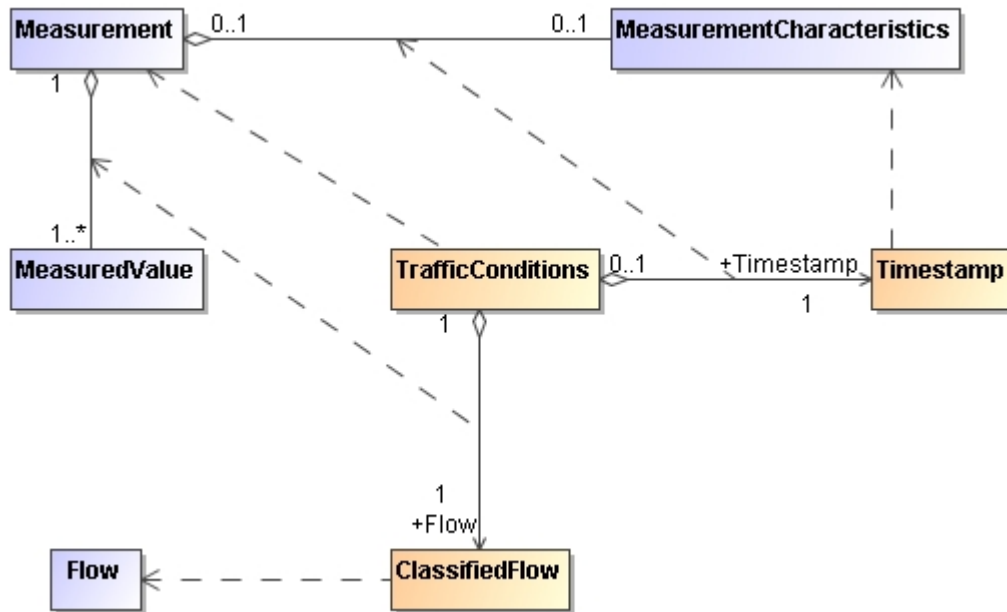


Figure 5 Mapping of entities and relationships

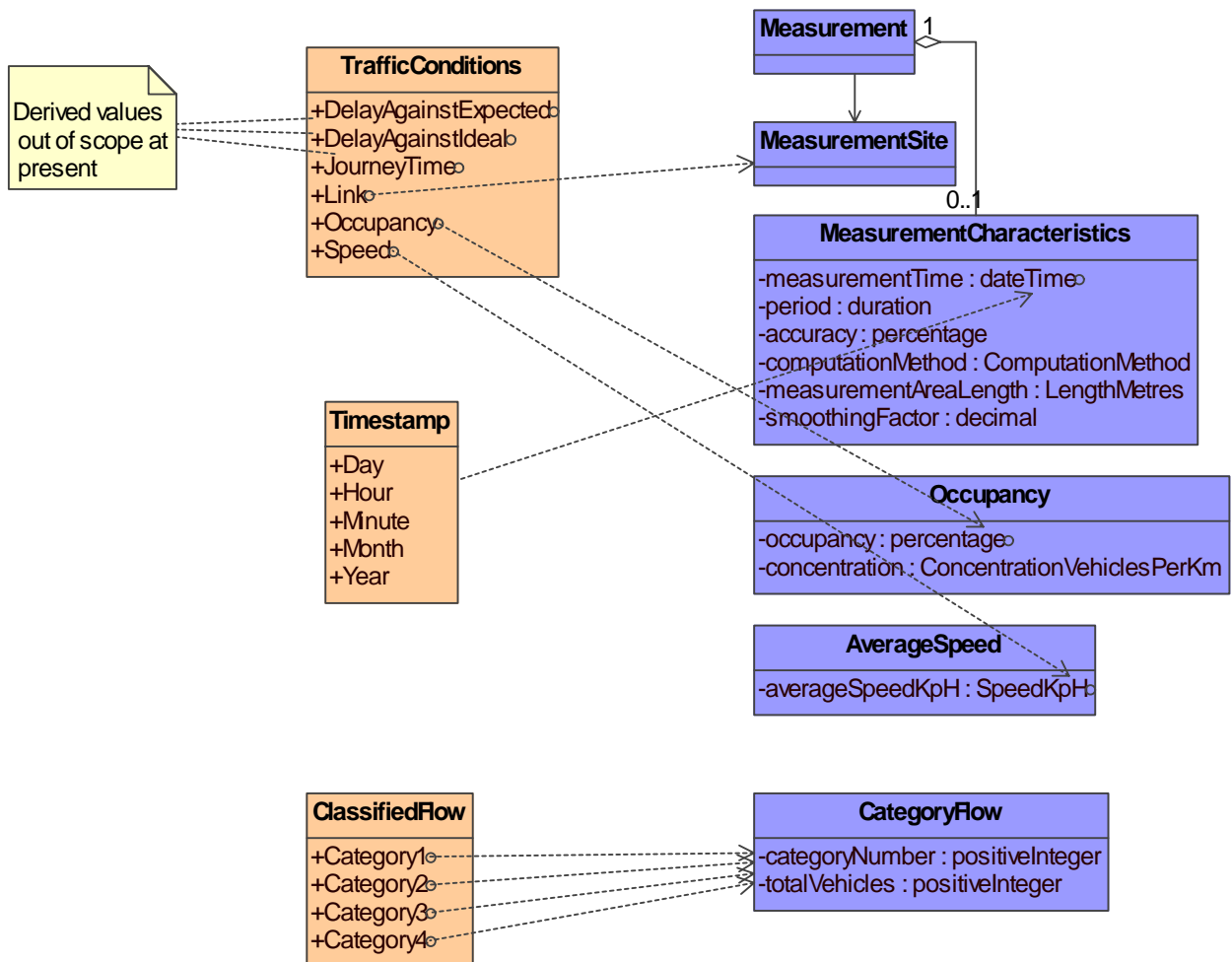


Figure 6 Mapping of attributes

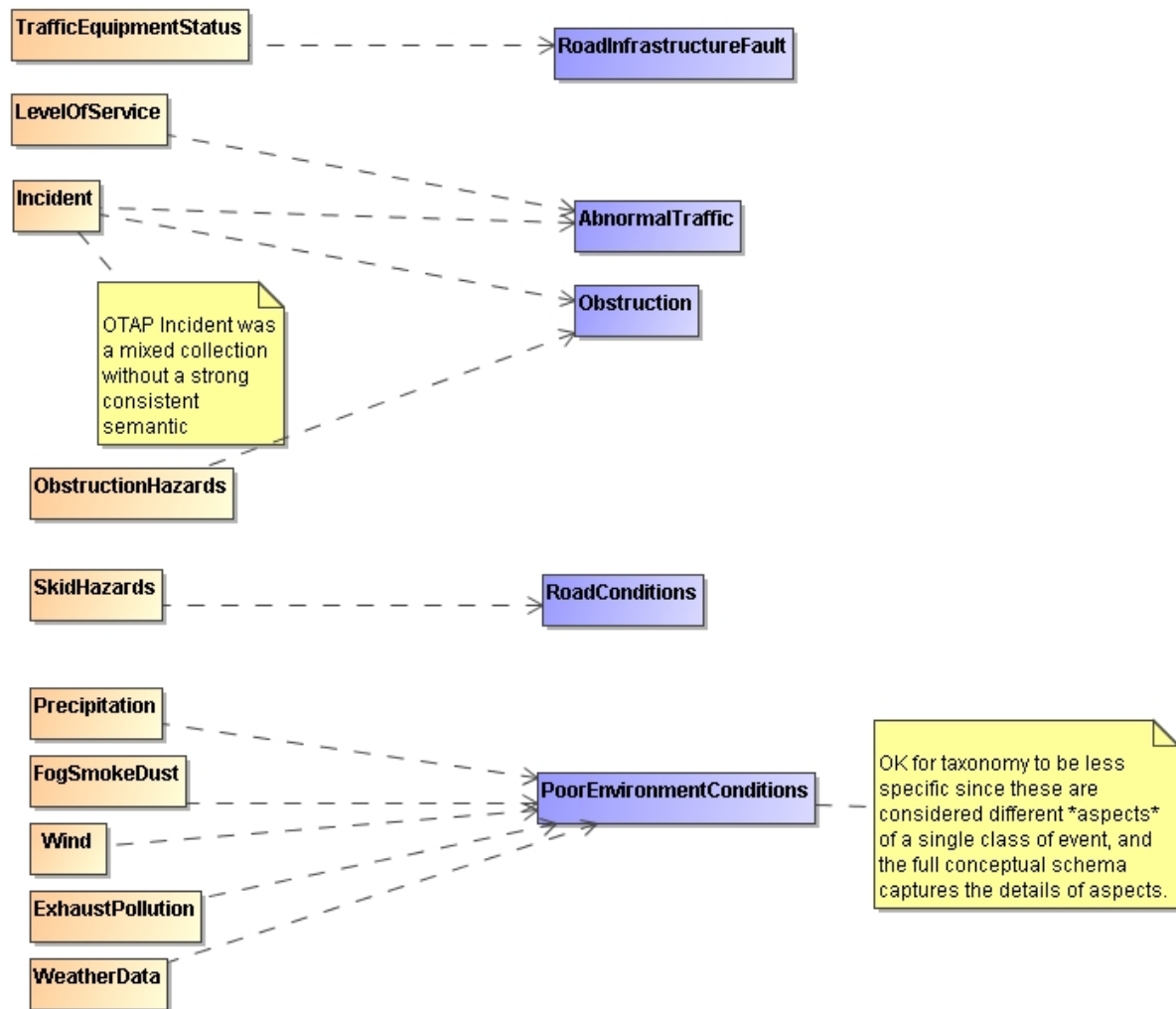


Figure 7 Mapping of taxonomy (fragment)

Each dependency should be given documentation (available in the registry), that states the business context, explains the differences to the core components, and if possible justifies these differences.

Submitted models in their original states may differ each other and from the core components:

- (i) for valid reasons of optimisation for their different business contexts
- (ii) in cases where there is no strong justification for the difference.

The end goal of the harmonisation process is for all differences of the second kind to be removed, and for the only remaining differences between models and core components to be clearly justified for the specific business context of those models. However, if no immediate change is possible then the understanding communicated by the core components and mappings is still valuable.

4 Conclusion

The Highways Agency core components analysis process evolved from the UN/CEFACT Core Components Technical Specification independently from the TBG 17 harmonisation team process, but has produced some similar guidance. However, the Highways Agency process is more focussed on

incremental improvements to legacy systems and is cautious about investment in new design. It is particularly suited to a context where there the harmonisation team does not have a strong mandate for change to existing systems.

The advantages of the core components analysis process are:

- It makes explicit the similarities and differences between existing specifications with overlapping semantics.
- The mappings process distinguishes justified design from flawed design.
- It generates objective feedback to submitters.
- The understanding can be used when building translators.
- It can be used within a registry process to identify candidates for recommendations (or “preferred” status), awarded in a specific business context.
- All the thinking is exposed to future designers

References

1. ISO/TS 15000-5:2005: Electronic Business Extensible Markup Language (ebXML) -- Part 5: ebXML Core Components Technical Specification, Version 2.01(ebCCTS).
2. UN/CEFACT TBG17 “BP&CC – Harmonization Team Submission Guidelines and Procedures.” Version 1.00 29 September 2004.
3. “Extreme Programming explained”, Kent Beck, Addison-Wesley 2000
4. “An eXtreme method for developing lightweight ontologies”, M. Hristozova, L. Sterling, in “Proc. of the OAS’02 Workshop”, 2002.