Position Paper: Modularity, Namespaces and Versioning

Author: Bill Burcham (bill_burcham@stercomm.com)

Date: 5/21/02

Filename: draft-burcham-modnamver-04.doc

1 Summary

There are many possible mappings of XML schema constructs to namespaces and to operating system files. This paper explores some of those alternatives and sets forth some rules governing that mapping in UBL.

2 Problem Description

Namespaces are a syntactic convenience supporting the association of a "context" with either a lexical scope (default namespace), or a shorthand identifier (namespace qualifier). This context, applied either implicitly (in a lexical scope) or explicitly (via qualified names) supports compression of what would otherwise be long identifiers. In the absence of namespaces, identifier names are all long.

It is common for an instance document to carry namespace declarations, so that it might be validated. Processing logic (such as a stylesheet) typically carries namespace declarations pertaining to the language in which it is specified in (XSLT) as well as the namespaces on which it *operates*. The latter must match namespaces in the instance document under translation in order for useful work to be carried out.

In practice, namespaces are often given names denoting a hierarchy. XML processing tools may or may not use this hierarchy information. This sort of hierarchical naming though can be useful for the human reader.

As with other significant software artifacts, schemas can become large. In addition to the logical taming of complexity that namespaces provide, we might like to also divide the physical realization of that schema into multiple operating system files.

Schemas change over time. UBL will be no exception. What sort of version information (if any) will a schema carry? How shall that information be carried so as to conveniently support the needs of users operating on document instances with XML processing tools.

This position paper will address these three topics related to namespaces:

1. **Namespace Structure**: What shall be the mapping between namespaces and XML Schema constructs (e.g. type definitions)?
2. **Module Structure**: What shall be the mapping between namespaces and XML Schema constructs and operating system files?
3. **Versioning**: What support for versioning of schema shall be provided?

In subsequent sections, we'll examine each topic in turn, presenting first the options, then a recommendation.

3 Assumptions

Much of this discussion will be based on the expected complexity of the UBL vocabulary. We structure systems into components in order to manage complexity.

3.1 Problem Size

How big will UBL be? How interconnected?

One source for complexity estimation is xCBL. TBD: how many type definitions, element declarations, "instance roots" in xCBL?

Another source for estimation is X12 that according to [NDR-MSG-88] has:

> a bit over 1,000 data elements (…) a smaller number of segments, and 300 or so transaction sets

Also from [NDR-MSG-88] we have EDIFACT:

- There are just under 650 data elements which are
- used in approx 200 composite structures (sort of equivalent to low level Aggregate Core Components (ACCs)).
- These elements and composites are reused within just over 150 segment structures (sort of equivalent to higher level ACCs).
- Combinations of all the above make up just under 200 messages (doc types).

So an estimate of 1000 types and 250 message types seems reasonable for UBL.

3.2 Optimal Component Size

We don't want to define 1000 types all in one XML namespace, nor would we want to define them all in one file. Such an approach would lack structure necessary for understanding both by maintainer and users. Additionally, performance would be far from optimal for instance documents that needed only a subset of the UBL types.

For these reasons we presume that we need to structure and divide UBL into a hierarchy of components. We will strive to balance coupling and cohesion between the components in order to:

- Manage the complexity of each component while not creating too many components[1]
- Provide for useful subsetting of components

We envision that many useful instance documents (messages) will be possible that require only a fraction of the overall UBL schema. In those cases it should be possible to avoid processing of the unneeded parts.

4 Options: XML Namespace Identification

This section presents some options for the form that UBL namespace names might take.

---

[1] The "seven plus or minus two" rule [SEVEN-TWO] is a good, general rule of thumb. It's especially useful when you don't have any other rule. It says that if you want people to be able to keep a set of concepts in mind, then you are limited to about seven concepts. Implications for XML for example might be: a type would define no more than seven (or so) elements, a namespace would define no more than about seven types, etc.

4.1 Option 1: Namespace Name = Namespace Location

There is certainly precedent for this approach. See for example the ebXML Message Service schema http://www.oasis-open.org/committees/ebxml-msg/schema/msg-header-2_0.xsd.

4.2 Option 2: Namespace Name is OASIS URN namespace

This option exemplifies the current best practice within OASIS. See RFC 3121 [OASIS-URN-NS] for details. See Namespaces in XML for background [NAMESPACE].

Under this option, the namespace names for UBL namespaces would have the following form while the schemas are at draft status:

`urn:oasis:names:tc:ubl:schema{:subtype}?:{document-id}`

When they move to specification status the form will change to:

`urn:oasis:names:specification:ubl:schema{:subtype}?:{document-id}`

Where the form of {document-id} is TBD but should match the schema module name (see section 6, Recommendation: Schema Location).

5 Recommendation: Namespace Identification

We pick Option 2: *Namespace Name is OASIS URN namespace.*

This recommendation probably needs more justification.

Will document-id include versioning information or will versioning be handled outside this identifier? See section 12, Recommendations: Versioning.

6 Recommendation: Schema Location

A question related to Namespace identification is schemaLocation. Schema location includes the complete URI which is used to identify schema modules.

In the fashion of other OASIS specifications, UBL schema modules will be located under the UBL committee directory:

http://www.oasis-open.org/committees/ubl/schema/<*schema-mod-name*>.xsd

TBD does this recommendation need more justification?

Where <schema-mod-name> is the name of the schema module file. The form of that name is TBD.

7 Options: Namespace Structure

In this section we'll explore some mappings between XML Schema structures and namespaces.

7.1 Option 1: One Big Namespace

We could have one big namespace for UBL. On the plus side, it would be fairly easy to remember. The downside is that we would forfeit the opportunity to use hierarchical namespaces to communicate the structure of the vocabulary.

7.2 Option 2: One Namespace Per Type

This approach represents the other end of the spectrum. If you've got a namespace per type then why not just use the type name. The namespace fails to be shorthand for anything. It fails to be memorable, or to group related types together.

7.3 Option 3: Core Plus "Functional" Namespaces

This option represents a space between 7.1 and 7.2. There would be namespaces for "core" types and there would be namespaces for each of the TBD functional areas e.g. Order, Invoice.

Purpose	**Namespace name**
Common Leaf Types	`urn:oasis:names:tc:ubl:CommonLeafTypes[TBD version info]`
Common Aggregate Types	`urn:oasis:names:tc:ubl:CommonAggregateTypes[TBD version info]`
Order Domain	`urn:oasis:names:tc:ubl:Order[TBD version info]`
Invoice Domain	`urn:oasis:names:tc:ubl:Invoice[TBD version info]`
TBD	TBD

This represents a top-level decomposition of the vocabulary into multiple vertical (functional) slices and a single (horizontal) slice – the so-called core.

The downside of this approach is that with seven or so functional namespaces, they are going to get awfully "crowded" (on the order of one hundred types per namespace).

7.4 Option 4: Core Plus "Functional" Namespaces Plus Internal Structure as Needed

A refinement on 7.3 this option frees each of the functional and core namespaces to have their own hierarchy as necessary in order to further manage complexity.

8 Recommendation: Namespace Structure

	Pro	Con
Option 1: one big namespace	Easy to remember namespace	When anything in UBL changes, all processing code must be changed (at a minimum to use new namespace name)
Option 2: namespace per type	Total compartmentalization	Why use namespaces at all? With this option the namespace

		With this option the namespace ceases to provide useful contextualization.
Option 3: core plus "functional" namespaces	Allows parts of UBL to change independently. When a functional area changes, processing code depending on core needn't change.	Doesn't allow for intermediate structure. What if the functional namespaces may require further subdivision?
Option 4: core plus "functional" namespaces plus internal structure as needed	(same as Option 3)	By allowing intermediate namespaces, they will certainly flourish. Design rules must be developed to avoid regressing toward Option 2 over time.

Option 3 is recommended. We reserve the right to revisit this decision when we are further along in the process of defining types. If we find that we need more structure, we can move to option 4.

8.1 Into What Namespace Do Extensions Go

Extensions (by users) go into user-defined namespaces outside of UBL.

9 Options: Module Structure

TBD: what are some other options?

10 Recommendation: Module Structure

This section describes the mapping of namespaces (as discussed in section 7 *Options: Namespace Structure*) onto files. A namespace contains type definitions and element declarations. Any file containing type definitions and element declarations is called a SchemaModule.

Every namespace has a special SchemaModule, a RootSchema. Other namespaces dependent upon type definitions or element declaration defined in that namespace import the RootSchema and only the RootSchema.

If a namespace is small enough then it can be completely specified within the RootSchema. For larger namespaces, more SchemaModules may be defined – call these InternalModules. The RootSchema for that namespace may then include those InternalModules.

This structure provides encapsulation of namespace implementations. To recap:

Import Rule: A namespace "A" dependent upon type definitions or element declaration defined in another namespace "B" imports B's RootSchema. "A" never imports other (internal) schema modules of "B".

Include Rule: The only place XSD "include" is used is within a RootSchema. When a namespace gets large, its type definitions and element declarations may be split into multiple SchemaModules (called InternalModules) and included by the RootSchema for that namespace.

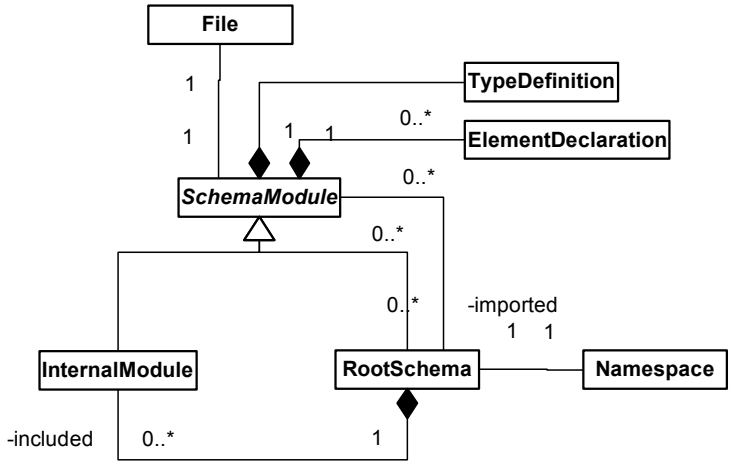
The preceeding rules present a namespace is an indivisible grouping of types. A "piece" of a namespace can never be used without all it's pieces. It is therefore important to strive to define namespaces that are minimal and orthogonal.

It is not enough that a namespace be minimal in terms of its intrinsic size, but also in terms of the closure of all other namespaces it imports. By closure we mean namespaces it imports, and namespaces they import, and so on.

One good way to foster minimal namespaces is to dictate that there be no circular dependencies between them. The same statement can be made for SchemaModules. This rule has been applied successfully in many large systems[2].

(No) Circular Dependency Rule: There are no circular dependencies between SchemaModules. By extension, there are no circular dependencies between namespaces. This rule is not limited to *direct* dependencies – transitive dependencies must be taken into account.

Here is a depiction of the component structure we've described so far. This is a UML Static Structure Diagram. It uses classes and associations to depict the various concepts we've been discussing:



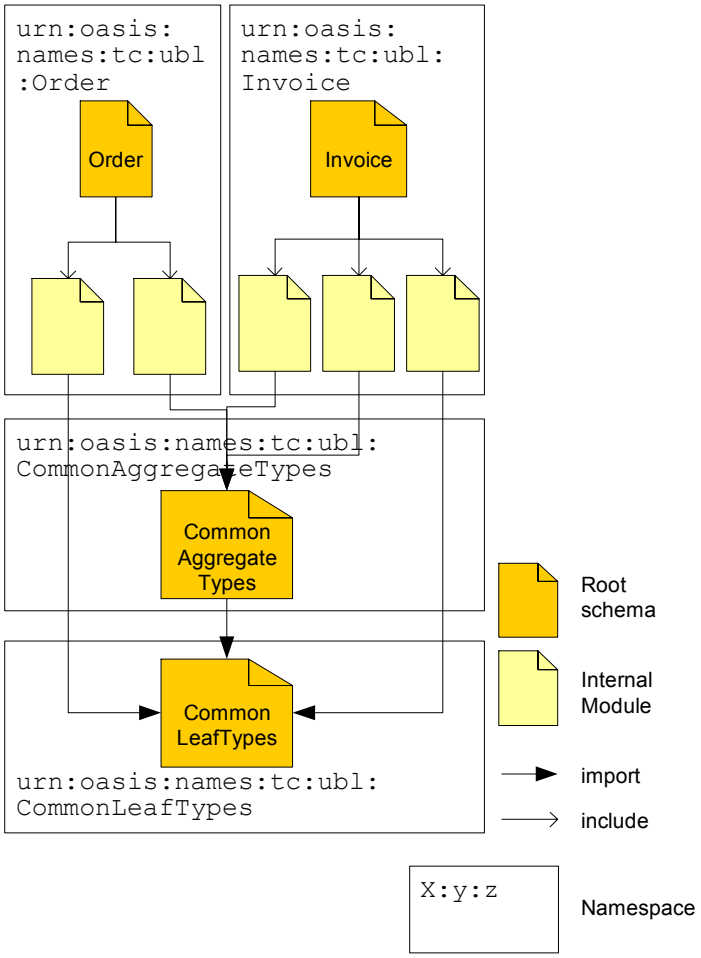
[2] For example [LARGE-SCALE] introduces the concept of "levelization" as an organizing principal for very large C++ systems. Those systems, due to the nature of the language often have an explosion of type definitions (due to the presence of parameterized types). As a result, solutions to the management of type systems in C++ could be viewed as exemplary for our purposes.

You can see that there are two kinds of schema module: RootSchema and "InternalModule". A RootSchema may have zero or more InternalModules that it includes. Any SchemaModule, be it a RootSchema or an InternalModule may import other RootSchemas.

The diagram shows the 1-1 correspondence between RootSchemas and namespaces. It also shows the 1-1 correspondence between files and SchemaModules. A SchemaModule consists of type definitions and element declarations.

The diagram unfortunately fails to express the (No) Circular Dependency Rule.

Another way to visualize the structure is by example. The following informal diagram depicts instances of the various classes from the previous diagram.



The preceeding diagram shows how the order and invoice RootSchemas import the "CommonAggregateTypes" and "CommonLeaf Types" RootSchemas. It also shows how e.g. the order RootSchema includes various InternalModules – modules local to that namespace. The clear boxes show how the various SchemaModules are grouped into namespaces.

10.1 Message Types

If preferring type definitions over global element definitions is good, why not take it to the extreme [NDR-MSG-70]. **The type of the root element of a UBL document (message) is a global type** (not an anonymous type).

10.2 Number of Message Types

In some cases, various actions in the protocol (create vs. delete) will have totally different document structure requirements. But in some cases (create vs. update), the content might be identical. However, we still think we should design in favor of more document types rather than less, e.g. one for each transmission (a la RosettaNet). It avoids confusion on the part of developers to have a separate document type for each thing. We might then decide to optimize some of them by merging them together.

11 Options: Versioning

[XFRNT-VER] does a great job of laying out the problem and solution space for schema versioning as it is traditionally practiced. The options presented in that document are not really disjoint rather they are building blocks. If you look at the recommendations in that document, you will see that the options are used in concert.

11.1 Option XF-1: Change the (internal) schema "version" attribute

11.2 Option XF-2: Create a "schemaVersion" attribute on the root element

11.2.1 Usage A: Conformance enforced by validator

11.2.2 Usage B: Conformance enforced by an extra processing pass

11.3 Option XF-3: Change the schema's target namespace

11.4 Option XF-4: Change the name/location of the schema

11.5 Option 5: Schema Version as Context Classifier

In [NDR-MSG-13] the point was made that schema version might just be another context classifier.

12 Recommendations: Versioning

Each namespace should have a version. Other things shouldn't (e.g. schema modules shouldn't).

Each of core and functional areas will have a version. How shall we communicate compatibility/incompatibility?

One approach is to follow a convention whereby a schema's version identifier consists of two parts: a major number and a minor number. The major number changes when a backward-incompatible change is made:

- changing a default value (legal issue)
- adding a new required element
- removing a required or optional element

The minor number changes when a backward-compatible change is made:

- adding an optional element

How do we communicate version compatibility between core and functional areas:

- between core and f/a's
- between f/a's

TBD: Include input from SAML versioning paper (to be released 1-10-02)

The following table summarizes the tradeoffs between the options.

	Pro	Con
Option XF-1: Change the (internal) schema "version" attribute		Not enforced by validator
Option XF-2-A: Create a "schemaVersion" attribute on the root element -- Conformance enforced by validator		Conformance requires exact match on version string
Option XF-2-B: Create a "schemaVersion" attribute on the root element -- Conformance enforced by an extra processing pass		Extra processing step.
Option XF-3: Change the schema's target namespace		With this approach, instance documents will not validate until they are changed to

		designate the new targetNamepsace. However, one does not want to force all instance documents to change, even if the change to the schema is really minor and would not impact an instance. +Include problems.
Option XF-4: Change the name/location of the schema		Ugh!
Option 5: Schema Version as Context Classifier	Leverages the context machinery	Requires the context machinery

13 Definitions

Backward compatibility – TBD.

BIE – Business Information Entity. A description of a business concept. Represented as an XML schema by a *root schema.*

extension a.k.a. customization – specification of new BIE's with well-defined, enforced relationships to old BIE's. Relationship types include: restriction, extension. In some cases processing logic will need to treat the base and the extension as the same, in other cases it will need to distinguish between them.

Forward compatibility – TBD

Namespace – a name that scopes a related group of XML type definitions.

processing logic – software logic that operates on BIE instances to achieve some business function

root schema – A *schema module* that directly, or via inclusion of other schema modules, defines all types for a particular namespace. This is the XML Schema representation of a BIE. (Compare that definition, with the one we came up with last week in Menlo Park: *A schema document corresponding to a single namespace, which is likely to pull in (by including or importing) schema modules*. **Issue**: Should a root schema always pull in the "meat" of the definitions for that namespace, regardless of how small it is?)

schema document – as defined by the XSD specification – per that specification, a schema document defines types into exactly one namespace, the target namespace.

schema module – A *schema document*. A schema module need not define all types in a particular namespace. Contrast with *root schema*. (Compare that definition, with last week's: *A "schema document" (as defined by the XSD spec) that is intended to be taken in combination with other such schema documents to be used.*)

versioning – reification of revisions to BIE's in order to support coexistence in a system, of two or more revisions of a BIE.

14 References

LARGE-SCALE	*Large Scale C++ Software Design*, John Lakos, 1996, Addison-Wesley.	
NAMESPACE	*Namespaces in XML*	http://www.w3.org/TR/REC-xml-names/
NDR-MSG-13	*schema version as context classifier*, Burcham, Bill; Maler, Eve; a post to the UBL-NDR mailing list.	http://lists.oasis-open.org/archives/ubl-ndrsc/200111/msg00013.html
NDR-MSG-70	*Fwd: Straw Man on Namespaces, Schema Module Architecture,etc.*, Rawlins, Mike; a post to the UBL-NDR mailing list.	http://lists.oasis-open.org/archives/ubl-ndrsc/200111/msg00070.html
NDR-MSG-88	*Fwd: Straw Man on Namespaces,Schema Module Architecture, etc.*, Probert, Sue; Maler, Eve.; a post to the UBL-NDR mailing list.	http://lists.oasis-open.org/archives/ubl-ndrsc/200111/msg00088.html
OASIS-URN-NS	IETF RFC 3121 *A URN Namespace for OASIS*	http://www.faqs.org/rfcs/rfc3121.html
SCHEMA-PRIM	*XML Schema Part 0: Primer*	http://www.w3.org/TR/xmlschema-0/
SEVEN-TWO	*The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information*, George A. Miller, Psychological Review, 63, 81-97.	http://psychclassics.yorku.ca/Miller/
XFRNT-VER	*XML Schema Versioning,* MITRE Corporation and xml-dev list group members.	http://www.xfront.com/Versioning.pdf
XML-URI-LIST	*XML-URI List* at w3.org	http://lists.w3.org/Archives/Public/xml-uri/