BII WG1 Controlled Vocabulary Approach Recommendations

G. Ken Crane Softwrights Ltd. Holman

Crane Softwrights Ltd.

<gkholman@CraneSoftwrights.com>

$Date: 2008/10/16 15:40:15 $(UTC) Published draft 4

Table of Contents

1. Introduction

This document proposes recommendations for Working Group 1 to make to the Business Interoperability Initiative [BII] with respect to the specification of controlled vocabularies and the related published artefacts for their support by users.

Typically, controlled vocabularies are used when specifying values for basic business information entities based on code lists or identifiers. Other basic entities may also be so specified when users wish to constrain the value used to be one of a controlled set of values.

Program-assisted data entry of controlled vocabulary entities ensures rogue values are not inadvertently specified for an information entity. Outboard document validation of controlled vocabulary entities ensures rogue values are not inadvertently passed to an application programmed to act on expected values. Such outboard processing can relieve the burden of implementing the same controlled vocabulary checking in all programs using documents, allowing the programs to work with supplied values on faith.

The BII must specify which entities are so constrained by the committee, the values by which these are constrained, and the associated semantics represented by the values. Interoperability is promoted by imposing these constraints, ensuring two parties interchanging documents use values that have accepted semantics.

Outside of the purview of the committee, users may themselves enforce further constraints on allowed values. Typically this will be by restricting the lists of applicable values to a smaller list of values. Such a subset list remains conformant for interoperability since all values of the smaller list are also values of the larger standardized list. Occasionally, users may wish to extend the lists of applicable values beyond the committee-specified values. Such a superset list is non-conformant for interoperability since there are no committee-wide agreed semantics published for these unanticipated values. Nevertheless, two trading partners may be compelled to break wider interoperability in order to achieve reciprocal interoperability by mutually agreeing on unpublished business-related semantics.

The BII must, therefore, publish artefacts representing committee definitions of controlled vocabularies such that users can apply them, restrict them or extend them as required.

2. Key terms and concepts

2.1. Controlled vocabulary

This document makes reference to certain terms related to controlled vocabularies. Such a vocabulary can be considered as a list of values where each value represents some agreed-upon semantic between parties using them. Codes and identifiers are often constrained by controlled vocabularies.

Using a controlled vocabulary to represent abstract concepts promotes internationalization. By associating multiple nuanced descriptions of the concept, users in different communities can better understand what others interpret through the values.

Controlled vocabularies mature and morph over time as different versions. Users determine some members no longer have any value in including, and other members are introduced due to needs of the community of users.

Some concepts might be best represented by a union values from more than one controlled vocabulary. Users may need to restrict a list to a subset or to augment a list to a superset in order to meet real-world business constraints and requirements. Sometimes they may need to augment a restricted list, thereby incorporating both concepts.

Supporting all of these facets of controlled vocabularies are three kinds of meta data to consider by the BII committee: value-level meta data, list-level meta data and instance-level meta data.

2.2. Codes and identifiers

Many of the information entities expressed in an exchange comprise totally arbitrary values without constraints on their representation. Consider text fields (e.g. "John Smith" as a name), amount fields (e.g. 123.45 units of currency) and numbers (quantity of 7 items). There is nothing abstract about such values.

Any information item in an exchange may, however, be constrained to only those values described in a controlled vocabulary. Each controlled value represents a concept that is commonly understood between parties, without needing to spell out the concept in possibly ambiguous words. Code lists and identifier lists are merely applications of controlled vocabularies, using values to represent abstract concepts or values to distinguish between members of a set.

Distinguishing that a particular item is represented by a code list or by an identifier list, that is distinguishing that the value is a code or an identifier, is often very arbitrary and sometimes quite unclear. For many the two concepts are indistinguishable. Is the value "`USD`" a code representing the abstract concept of a US dollar, or is it an identifier representing the US dollar currency? Both are meaningful statements and each beholder may see one interpretation or the other. Some would say identifiers are easily recognized such as product identifiers and service identifiers offered by an organization, though who hasn't heard of the term "product codes" for such information?

Those deploying information interchange will end up deciding to call some concepts "codes" and some concepts "identifiers" and their representation is then constrained to the expression modeled for such values. When the BII requirements are deployed in actual message definitions, most of the concepts constrained by controlled vocabularies will be called codes or identifiers. The BII will be responsible for publishing the values and their associated meanings as the controlled vocabularies.

Users of the information being interchanged may impose their own restrictions by constraining other arbitrary information in BII documents. These users may cite BII vocabularies, other published vocabularies, or they may publish their own controlled vocabularies comprised of values and their associated meanings, for use between trading partners.

The rest of this document refers primarily to "controlled vocabularies" without distinguishing very often between code lists and identifier lists. Those who deploy BII requirements as document definitions and those who use these deployed documents will be able to constrain the values of any information item to controlled vocabularies using the concepts and facilities described in this document.

3. Controlled vocabulary challenges

A controlled vocabulary is distinguished from free-form entry of data by constraining the value of an information item to be one or more values from a set of prescribed values, proscribing any value not included in the set. The custodian of a controlled vocabulary dictates what each value represents so that all users of the value (hopefully) have a common understanding when using or encountering it. See Section 4.1, “Value-level meta data” regarding a formal method of documenting or describing properties of each value in a vocabulary.

All values in a controlled vocabulary must be unique though there is no obligation that the value must be different from values found in other controlled vocabularies. The meaning of a value used or encountered is, therefore, identified by the context of the value being in a particular list of values. Thus, the identity of a list of values is important to agreeing upon the semantics of values used from that list. See Section 4.2, “List-level meta data” regarding a formal method of documenting or describing properties of the vocabulary as a whole.

Often a value's meaning can be inferred as being from a particular list just by its context of use in a document, that is, the entity in which the value is found. Making this inference is based on the assumption that only a particular list is being used in that context without ambiguity. Where it is necessary for a single information entity to allow the same specified value to be used from two different lists (thus representing two different semantics), there is an obligation to disambiguate the meaning of the value by specifying from which list the value is being used. See Section 4.3, “Instance-level meta data” regarding a formal method of disambiguating values used in an interchange.

Values are typically, but not necessarily, abbreviations that are conveniently concise when used in message formats. A controlled vocabulary may, however, be a set of values of any length for any purpose, using any character set and any lexical representation rules (e.g. the presence or absence of spaces allowed in the value).

International organizations take it upon themselves to specify controlled vocabularies for common global concepts such as country codes or currency codes. All trading partners are obliged to interpret values in these lists the same way or interchange risks being ambiguous and failing. Meanwhile, businesses and trading partners will take it upon themselves to specify controlled vocabularies for concepts under their own purview such as product and service identifiers. Who better would know what identifier represents a particular custom unique concept or item belonging to the business? A trading partner wishing to interact with the business must utilize these values or interchange will risks being ambiguous and failing.

3.1. Internationalization

Values in controlled vocabularies are either mnemonic in some fashion, or totally abstract if the concept represented cannot be somehow derived from the representation itself.

Mnemonics may be based on a single language for all values, or may be in different languages based on some context, provided that all values in the list are unique. There is no set rule regarding which mnemonics to use. Consider that in the international controlled vocabulary representing languages of the world "`EN`" represents English, "`ES`" represents Spanish (Español), but the choice to use Roman orthography and Western European character sets results in Japanese being represented by "`JA`" and not "日本語" (noting interestingly that "`NI`" would be available for "Nihongo" and yet was not used).

For private interchange and custom code lists there are no restrictions to use Roman orthography and Western European character sets. Japanese users of private lists could very well use Japanese orthography and Unicode in the values representing concepts. Doing so would understandably limit interchange with trading partners unfamiliar with the values or unable to decode the meanings represented by the values.

An example of a totally abstract representation is the international controlled vocabulary for financial payment means: there is no way to derive from the value "`42`" that it represents the concept of "payment to bank account".

That a single value of any kind is decided to represent a particular concept promotes the internationalization of documents. All users of all languages are obliged to use the commonly-understood fixed value when wanting to express or confirm an information item represents a particular concept. There is no ambiguity that would otherwise be introduced by users adopting their own abbreviation or representation for a concept.

Of course it is critically important that the values themselves be documented in such a way that international users of the values unambiguously understand what the value represents. See Section 4.1, “Value-level meta data” for a discussion of how important it is to associate supplemental

information to disambiguate the concepts behind values, and how facets of a value's description can incorporate multilingualism.

3.2. Versioning

Concepts come and go and thus controlled vocabularies change over time. Values representing defunct concepts are discarded from the list and values representing nascent concepts are introduced into the list. Interestingly, some custodians of controlled vocabularies re-use a value that once represented an original concept in one particular version of a list as representing a distinctly different concept in another version of the same list.

Sometimes a value's meaning might change over time because the concept itself has changed, rather than a conscious choice of the custodian to represent a different concept. Consider the country sub-entity representation "NT" of the Canadian province Northwest Territories. In 1999 Canada split what was Northwest Territories into two territories named "Northwest Territories" and "Nunavut". Indicating a region of Canada by its sub-entity representation "NT" now indicates different areas depending on whether the value is taken from a list of values dated before or after 1999. To unambiguously indicate which region a value represents requires one to simultaneously indicate from which list the value is taken. This might be done within the message itself or by some out-of-band mechanism.

Therefore, one needs to not only unambiguously identify a list of concepts as a controlled vocabulary, one may simultaneously need to be able to unambiguously identify a particular version of that vocabulary. See Section 4.2, “List-level meta data” for a discussion of disambiguating the identity of a list of values for a controlled vocabulary.

3.3. Restriction and extension

Users of a controlled vocabulary may find that there are too many or not enough concepts represented by the values therein. When there are too many concepts then the new constraints are a restriction of the available values. When there are too few concepts then the new constraints are an extension of the available values. There may in fact simultaneously be too many and too few concepts in the list requiring an extension of a restriction of the available values.

Perhaps the list is an international list and business constraints restrict the use of a subset of values (e.g. limiting the allowed currency values or limiting the allowed payment means). There is no attempt to redefine the meaning represented by any of the values in the restriction, the only objective is to limit which values can be used in an interchange.

The restricted list is, in fact, a different list than the original list, but the values in the restricted list unambiguously represent the concepts from the original list. The restricted list can, therefore, successfully masquerade as the original list where it is applied in documents, even though it has its own identity and versioning behind the mask.

Perhaps the list is an international list and business opportunities go beyond the available concepts represented by the values (e.g. there is no payment means representing barter). A value in the extension of the list represents, assumedly, a different concept than any represented in the original list. If it were represented then meaningful interchange obliges the use of the value in the original list. Meaningful interchange cannot, however, be guaranteed when representing an extended set of concepts for an information item. Only with a priori agreement between trading partners can successful interchange be anticipated. Blind interchange of an extension value cannot be expected to be interpreted as representing the same concept to both parties (though in practice often judicious choice of mnemonic extension values may logically lead to a common understanding).

The extension of the list is a different list than the original list. The extension of the list does not include the original list, as the original list is still under the aegis of its custodian. An information item governed by an extended set of values is, in fact, governed by a union of values from the original list and from the extension of the list.

This union of separate lists cleanly distinguishes the semantics or meanings behind the values in a way that creating a custom standalone list of all values does not. A custom list of all values cannot masquerade as the original list because there are members not in the original list. Moreover, the custom list cannot be applied in documents with its own identity and versioning because there are no universally-respected semantics associated with the list identity.

See Section 4.3, “Instance-level meta data” for a discussion of the use of information associated with a value in a document to establish the associated semantics when working with restricted and/or extended lists.

4. Meta data (the solution)

4.1. Value-level meta data

The BII committee needs to publish sufficient information for users to understand the semantics represented by a particular value to be used in interchange. This may include machine-readable facets and/or human-readable prose facets. Where a particular facet of a value's semantics is in prose, there may be an opportunity to express the semantic in different languages promoting a wider common understanding of meaning.

Each facet of definition can be considered to be value-level meta data.

Value-level meta data is responsible for conveying both identity and meaning. The lesser amount of such information available, the more opportunity for ambiguity. More such information promotes common understanding of the concept behind the representation in the value.

Requirements for internationalization can be met in value-level meta data by including multiple language entries for descriptions, as well as non-language components understood from the semantics of the concepts.

Many currently-published controlled vocabularies, such as ISO country codes, have very limited value-level meta data comprised only of the code and its name (where its name is meant to convey its definition).

For some controlled vocabularies, such as ISO location codes, only using the name does not disambiguate the definition of the associated value. More meta data information is required for each value to ensure two users are unambiguously agreeing upon the same location.

Figure 1. Controlled vocabulary value-level meta data

Code	Name	County	Subdivision	Port, rail, road, air, post, multi-modal, fixed, border	IATA	...
ADALV	*Andorra la Vella*	*Andorra*		*3,4,6*	*ALV*	...
...	...	...	...	...	...	...
USCB8	*Columbus*	*United States*	*MT*	*2,3*	*CB8*	...
USCBW	*Columbus*	*United States*	*WI*	*3*	*CBW*	...
USCLU	*Columbus*	*United States*	*IN*	*3,4*	*CLU*	...
USCMH	*Columbus*	*United States*	*OH*	*4*	*CMH*	...
USCSG	*Columbus*	*United States*	*GA*	*3,4*	*CSG*	...
USCUS	*Columbus*	*United States*	*NM*	*3,4,B*	*CUS*	...
USCZX	*Columbus*	*United States*	*NC*	*3,6*	*CZX*	...
USOLP	*Columbus*	*United States*	*MO*	*3,6*	*OLP*	...
USOLU	*Columbus*	*United States*	*NE*	*3,4*	*OLU*	...
USUBS	*Columbus*	*United States*	*MS*	*3,4*	*UBS*	...
USUCU	*Columbus*	*United States*	*KS*	*2,3*	*UCU*	...
USVCB	*Columbus*	*United States*	*TX*	*3*	*VCB*	...
USVDA	*Columbus*	*United States*	*MI*	*4*	*VDA*	...
USYBC	*Columbus*	*United States*	*NJ*	*2,3,6*	*YBC*	...
...	...	...	...	...	...	...
ZWWKI	*Hwange*	*Zibabwe*		*4*	*WKI*	...

"UN/ECE Rec 16 LOCODE"

Figure 1, “Controlled vocabulary value-level meta data” shows some of the rows and some of the columns from "UN/ECE Recommendation 16 LOCODE". Note how even using name and country is insufficient to disambiguate the location represented by a code, and that subdivision must also be used.

While none of the other fields of value-level meta data are needed for disambiguation, they are nevertheless very useful in conveying information to users to supplement meaning.

Requirements for versioning of individual values are also met in value-level meta data. Though list-level meta data reflects the version of the entire list, facets of definition of individual entries may change from list version to list version. By defining value-level meta data users custodians can augment outboard documentation with rich inboard information.

When creating one's own restricted lists from public lists it is important to copy all of the value-level meta data for the values being used so that applications do not lose fidelity of the description created by the custodians of the list. While a restricted list's list-level meta data can be masqueraded as the complete list, each list's entry is the only place in which to find the supplemental information for that entry.

When creating one's own extended list to augment another list, it will probably help users to find as much value-level meta data in the extended information as is found in the original list. This will allow applications to access all value-level meta data consistently without having to handle special cases for extended values.

4.2. List-level meta data

The BII committee needs to identify the lists from which values are found to ensure users associate the appropriate value-level meta data to a given value. Every published controlled vocabulary must, therefore, be uniquely identified with list-level meta data.

Not many custodians of internationally-published lists of values have managed the list-level meta data associated with their publications. Where such meta data is underspecified it becomes the responsibility of groups like the BII committee to use unsanctioned values of list-level meta data to meet the needs of users.

Figure 2. Controlled vocabulary list-level meta data

```
ShortName                    = PortCode
LongName (xml:lang="en")     = Port
LongName                     = UN/ECE rec 16
(Identifier="listID")
Version                      = 2006-2
CanonicalUri                 = urn:oasis:names:specification:ubl:codelist:gc:PortCode
CanonicalVersionUri          = urn:oasis:names:specification:ubl:codelist:gc:PortCode-2.0-upda
LocationUri                  = http://docs.oasis-open.org/ubl/os-UBL-2.0-update/cl/gc/special-
                               purpose/PortCode-2.0.gc
Agency                         LongName                = United Nations Economic Commission fo
                               (xml:lang="en")           Europe
                               Identifier              = 6
```

Figure 2, “Controlled vocabulary list-level meta data” is a snapshot of a rendering of list-level meta data for "UN/ECE Recommendation 16 LOCODE" as published for the Universal Business Language [UBL 2.0]. Note how the location URI is pointing to the UBL deliverable rather than a more preferred home at UN/ECE or UN/CEFACT. This is anticipated to change when UN/ECE or UN/CEFACT publish their lists in a machine processed form.

The fields shown are defined in genericode (see Section 6.1, “Genericode files”) as follows:

- `ShortName` (mandatory)
  - a reference to the list, without spaces, for use in referencing in programming
- `LongName xml:lang=` (optional and repeatable)
  - a prose name for the list, which may contain spaces, identifying the list in a particular language
  - as many long names can be included as there are languages to describe the list
  - in the absence of `xml:lang=` there are no presumptions about the language of the name of the list
- `LongName Identifier="listID"` (optional)
  - a language-neutral identifier for the list for use in referencing in programming
- `Version` (mandatory)
  - a label distinguishing this version of the list from other versions of the same list
- `CanonicalUri` (mandatory)
  - a universal resource identifier for the list independent of any particular version of the list, thus representing all possible versions of the list
- `CanonicalVersionUri` (mandatory)
  - a universal resource identifier for this particular version of the list, distinct from all other versions of the list
- `LocationUri` (optional and repeatable)
  - a pointer to a reference copy of this particular version of the list in genericode format
- `AlternateFormatLocationUri` (optional and repeatable)
  - a pointer to a copy of this particular version of the list in some non-genericode format
- `Agency` (optional)
  - information regarding the custodial agency for this controlled vocabulary
    - `ShortName` (optional)
      - a reference to the agency, without spaces, for use in referencing in programming

- `LongName` (optional and repeatable)
  - a prose name for the agency
- `Identifier` (optional and repeatable)
  - an identifier for the agency as assigned by some authority

It is important to note that any subset of a published list is obliged to have different list-level meta data than the list from which the values are obtained. This is because, in fact, the subset list is not the same list and so should not ambiguously claim to be the same list. However, the semantics of the items in the subset list are identical to the semantics of the items in the complete list. Thus, it will be necessary to be able to masquerade the subset list as the complete list such that the semantics of the subset list values are interpreted from the definitions found in the complete list.

Internationalization is addressed in list-level meta data by using a number of `LongName` elements in different languages. The language-neutral identifier works across all languages and version and resource identifier information values are all language independent.

When creating restricted lists, users are obliged to divine their own list-level meta data for their list as their list is, necessarily, different than the original list. It is unacceptable to re-use the meta data of the original list in the restricted list as the restricted list is not one created by the custodians of the original list. However, applications expecting list-level meta data from the original list will not recognize the list-level meta data from the restricted list, so when the restricted list is deployed in any application setting it is necessary to masquerade the restricted list as being the original list. See Section 7.2, “Context/value association files” for an example of how a restricted list can be masqueraded as the original list from which it was derived.

When creating extended lists, users are obliged to divine their own list-level meta data for that list that contains the selection of members not in the original list. Using a value that is from an extended list is constrained by the union of a base list (set of original values or restricted subset of original values) and the list extension (extended values).

See Section 4.3, “Instance-level meta data” regarding how one can distinguish the applicable list-level meta data for a value in a document through that value's associated instance-level meta data.

4.3. Instance-level meta data

The BII committee needs to specify mechanisms by which an information entity's controlled vocabulary value can be unambiguously interpreted. Typically this is done by supplemental information associated with the value indicating the list-level meta data for the controlled vocabulary from which the value is obtained. Instance-level meta data is that information added to the instance, that is, the values of list-level meta data associated with the controlled vocabulary from which the value is obtained.

Such instance-level meta data is typically optional as most entities allow values from only a single controlled vocabulary and values are unique within any given vocabulary. However, such meta data should always be available to be used since the BII will never know what business requirements users may need to satisfy through using custom lists, extensions of lists or unions of different lists for any code or identifier information entity.

Typically instance-level meta data refers to the language-neutral facets of identifying a controlled vocabulary, though designers may wish to offer the specification of a prose name.

Figure 3. Controlled vocabulary instance-level meta data

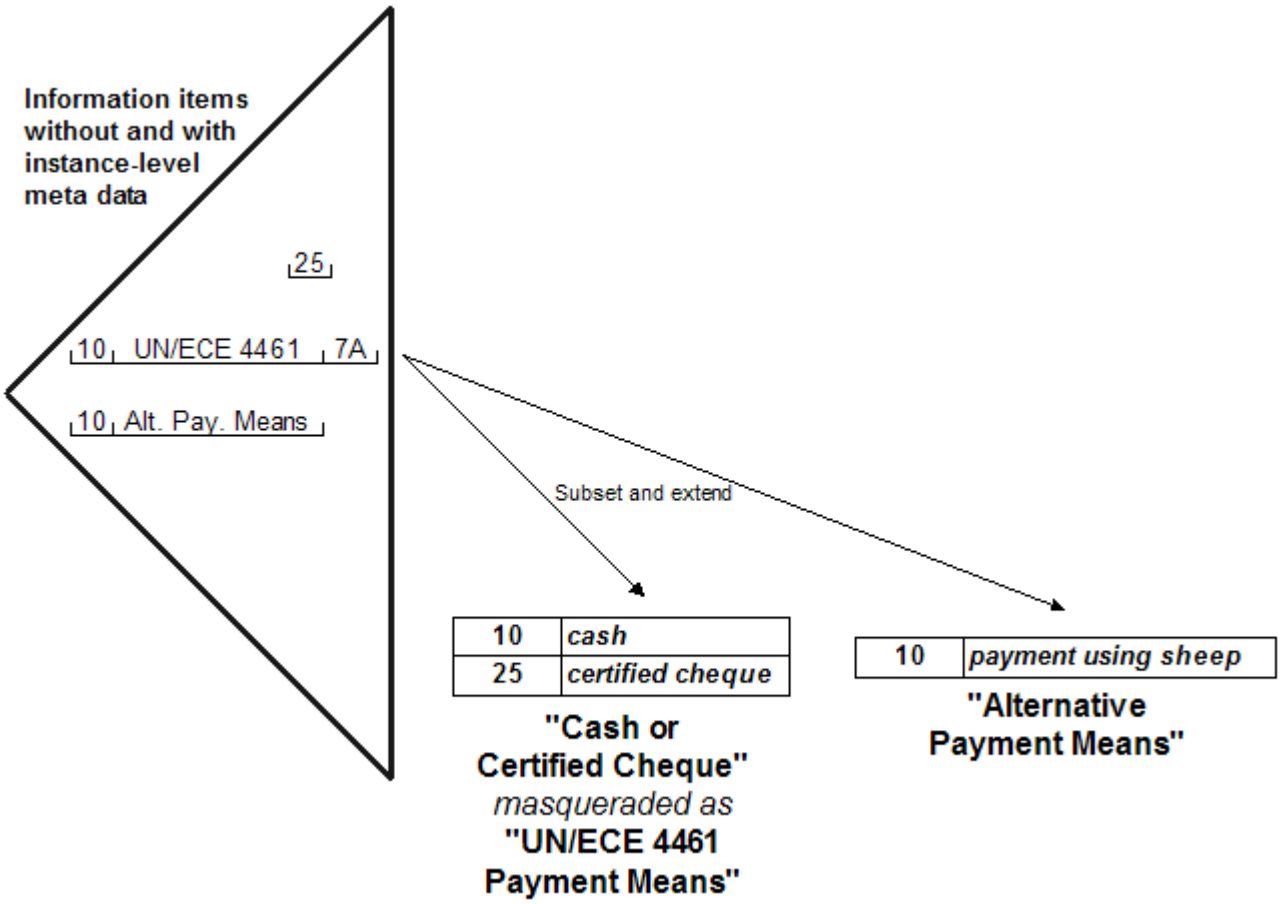


Figure 3, “Controlled vocabulary instance-level meta data” depicts an example of three information entities found in an XML document where codes are being used to identify the means of payment. For the scope of all three entities, a subset of the "UN/ECE 4461 Payment Means" list is extended in union with a non-conformant extension called the "Alternative Payment Means" list. Note that the title of the subset list is "Cash or Certified Cheque" since it cannot be named identically to the UN/ECE list (as it is, of course, not the list but a subset). However, the meaning of the values are defined by UN/ECE and so the subset list is said to be masquerading as the complete list.

The value "`25`" is being used in the first information entity. This value is unambiguously from the UN/ECE list. The second and third information entities are using the value "`10`". Alone, each value is ambiguous as it could represent a value from either the subset list or the extension. Through the use of instance-level meta data the first of these unambiguously represents the "cash" semantic from the UN/ECE list and the second unambiguously represents the "payment using sheep" semantic from the alternative list.

Instance-level meta data is critical to supporting versions of lists, custom lists and extended lists, especially where values in these lists are ambiguous with values in publicly-available lists. Without such meta data available, a naked value in a document cannot be accurately interpreted by an application making assumptions about the controlled vocabulary that applies.

Indeed one can choose, and often does choose, not to include instance-level meta data for publicly-available lists in expected document contexts, presuming the application is going to "do the right thing" with the naked value. But users requirements are unexpected and applications need the supplemental information identifying controlled vocabularies when users need to specify unexpected or ambiguous values representing concepts not assumed by the application.

5. Committee-recommended values for entities

Working group one is responsible for recommending to the BII which information entities, such as codes or identifiers, will have controlled vocabularies for interoperability. It will recommend which lists will be used for each controlled vocabulary, and the custodian of the semantics represented by the list values. When the BII needs to extend a controlled vocabulary with custom values it will indicate where multiple lists are in union for a single information entity, comprised of the original list (or subset thereof) and the extended BII list. It will describe the versioning technique and the impact on users.

This "approaches" document will not enumerate such recommended lists. These will be found in the document tentatively titled "BII WG1 Controlled Vocabulary Content Recommendations".

5.1. Values published by others

Working group one will be recommending, where available, the semantics defined by custodians of existing publicly-available controlled vocabularies. Organizations such as UN/ECE, UN/CEFACT and ISO are the custodians for many vocabularies in global use. All identifying list-level meta data will be reported for use in instance-level meta data.

Where the BII wishes to specify only a restricted subset of another controlled vocabulary, it will need to indicate which list-level meta data is to be used to identify the subset, and which list-level meta data represents the original set from which the subset is derived.

5.2. Values published by the BII

Working group one will be publishing, in the role of being the custodian, the semantics defined for values in controlled vocabularies created under the BII purview. As much machine-processed information will be included in value-level meta data.

For internationalization, where an item of value-level meta data is in prose, as many languages will be made available as is practicable or meaningful for the value. Using a complex meta data description value with `xml:lang=` attributes, one will be able to find a particular language's version of the description.

Where the BII wishes to specify an extension of another controlled vocabulary, it will need to specify which list-level meta data is to be used to identify the extension, and which controlled vocabulary's values are to be in the union of values applicable.

5.3. Controlled vocabulary versioning technique

All controlled vocabularies under BII purview will specify list-level meta data identifying the vocabulary independent of any particular version as well as list-level meta data identify the particular version of the vocabulary. In Figure 2, “Controlled vocabulary list-level meta data” such values are identified, respectively, as the canonical URI and the canonical version URI.

Similarly, the short name, long name, identifier and agency information for a given list will be the same list-level meta data value for all versions of the list, while the version information will differ for each version of the list.

By having both persistent and specific list-level meta data, values representing semantics not expected to change can be disambiguated using persistent meta data. Values representing transient semantics or changing semantics can be disambiguated using specific meta data ensuring a particular meaning is associated with the value.

The BII will need to choose how versions are represented in both the version indicator and the version universal resource identifiers.

5.4. Instance-level meta data

Working group one recommends the implementation of controlled vocabularies for code list or identifier values include provisions for specifying instance-level meta data associated with values. These supplemental components of data values allow the users to specify as much or as little meta data for a value in an instance as they feel necessary.

Some strategies may lead to little use of meta data for as much flexibility in downstream processing and interpretation. Other strategies may require users to very precisely identify the semantics associated with a value by indicating detailed list-level meta data associated with the value.

6. Committee-published artefacts

Working group one is responsible for deciding and documenting what code list and identifier values are permitted in BII documents and where they are permitted. It will also decide and document which values for code lists and identifiers are not being constrained by BII, thereby allowing an infinite set of values in a valid BII document.

As the BII is not responsible for creating schemas themselves, these committee artefacts are independent of the XML vocabulary used to implement the BII requirements.

6.1. Genericode files

A genericode file is a standardized XML expression of a controlled vocabulary, thus it can be used to represent the sets of values constraining code lists and identifier lists. It a declarative file describing all of the values has no preconceived use or mandatory application. For example, data entry programs could use genericode files to constrain the user input of values into a message. As another example, validation programs could use genericode files to confirm values found in a message satisfy constraints imposed.

Working group one will find, where available, or create machine-readable lists of values expressed in the OASIS genericode standard [genericode] specified by the OASIS Code List Representation Technical Committee [CLRTC].

Where genericode files already exist working group one will point to where the latest version of the resource can be found.

Where genericode files do not already exist, working group will be creating them and making them available to implementers. This will require working group one to arrange for the long-term maintenance of those values and their associated meta data, specified and maintained by the BII.

6.2. Constrained code list and identifier values

Not all code list and identifier values in BII documents are expected to have their values constrained.

Working group one will recommend to BII which information entities are to be constrained and will indicate the list of genericode files expressing the corresponding constrained values.

No standard specification exists for expressing this correspondence independent of a particular application of an XML vocabulary of structured elements and their attributes. This information could be conveyed simply in, for example, a spreadsheet cell for an information item indicating the value lists applicable. Such a specification will be more complex if there are contextually different constraints, that is, the values applicable for a single information item differ for that information item found in different document contexts.

6.3. BII document business rules

Working group one will determine and document, where applicable, other document constraints unable to be expressed using schema constraint semantics. A basic example of this would be where the value of one information item is dependent on the value of another information item, otherwise known as a co-occurrence constraint.

No standard specification exists for expressing this correspondence independent of a particular application of an XML vocabulary of structured elements and their attributes. This information could be conveyed simply in prose in, for example, a spreadsheet cell for an information item indicating the rules applicable. Such a specification will be more complex if there are contextually different constraints, that is, the rules applicable for a single information item differ for that information item found in different document contexts.

7. Implementer-published artefacts

This document assumes that an implementer is implementing the BII requirements by creating the artefacts expressing the constraints on BII XML documents. The artefacts in this section are merely examples of what might be used in an implementation of BII XML documents.

It is assumed the BII committee will not constrain or direct implementation approaches beyond stating the requirements to be fulfilled. The examples in this section are included only for discussion purposes and not an indication that the BII will publish any particular artefact documented here. They illustrate examples of how implementations might take advantage of the requirements published by the committee.

7.1. Schema structural constraints

The structural constraints of elements, attributes and the lexical structure of values is typically expressed using schemas. Any schema language could be used, for example W3C Schema [W3C Schema] or RELAX-NG [RELAX-NG]. In the rest of this discussion although W3C Schema is referenced, in fact any validation technology could be replaced in the prose or diagrams.

7.2. Context/value association files

An implementer could create machine-readable lists of document contexts expressed in the OASIS Context/value association (CVA) file specified by the OASIS Code List Representation TC. This specification is under development and the latest version is available from the committee home page [CLRTC].

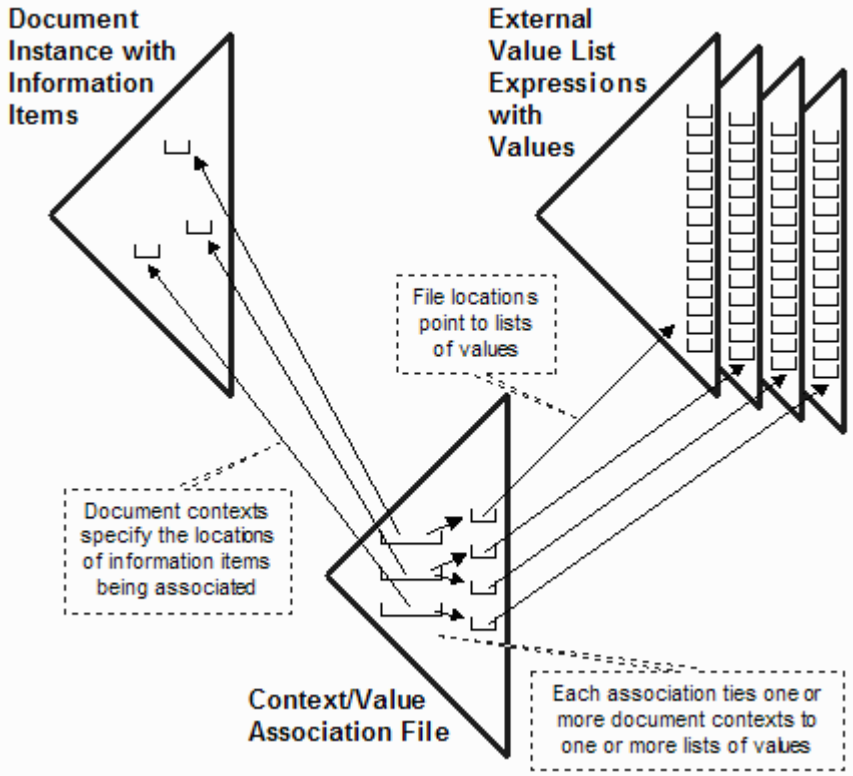
Using this XML document, developers of user interfaces and document data entry applications and services will be informed of which coded values in a document are permitted to have values from which code lists or identifier lists.

The document will also be useful in implementing validation processes used to confirm that constrained values do not violate their constraints. The constraints are the values in genericode files associated with the information item in document context.

Where necessary, a CVA file can specify that two different contexts of the same type of code list or identifier can be constrained by sets of different genericode files.

Figure 4, “Controlled vocabulary context ” illustrates the associations in a CVA file of external value lists with the document contexts of information entities in BII documents.

Figure 4. Controlled vocabulary context



7.3. Document business rules

Broadly defined, a document business rule is a constraint on document content that cannot be expressed (or easily expressed) in schema semantics or controlled vocabulary semantics. An example might be the association between a mandatory element and an optional attribute. This simple relationship is easily expressed in schema semantics. If, however, the optional attribute becomes a mandatory attribute in the presence of a particular value for the mandatory element (say a value greater than 10,000), there are no schema semantics to express this as a document constraint. A business rule can, however, precisely express the constraints on content values in the presence of other content values, known as co-occurrence constraints.

The BII will need to document what are the business rules of BII documents.

Implementers will need to express the BII document business rules in terms of their implementation of BII information entities in their document structures.

The ISO/IEC 19757-3 Schematron [Schematron] specification can be used to formally express non-schema BII document constraints in a machine-processed form for a particular XML vocabulary.

8. End-user options

As all of the BII resources created by working group one are declarative in nature, it will be up to implementers to create the artefacts used by end users. End users, in turn, employ these artefacts in their data flows and processes.

The BII will not be recommending implementers and end users utilize any particular processes, though there are some already publicly available that can be employed for some tasks. The examples in this section are included only for discussion purposes and not an indication that the BII will

publish any particular artefact documented here. They illustrate examples of how implementations might take advantage of the requirements published by the committee.

For example, Figure 5, “Simple two-pass validation of structure and value constraints” shows the use of a W3C schema expression (labeled "1") for the validation of the structure of elements and values, combined with the use of an XSLT expression (labeled "2") for the validation of values themselves.

Figure 5. Simple two-pass validation of structure and value constraints

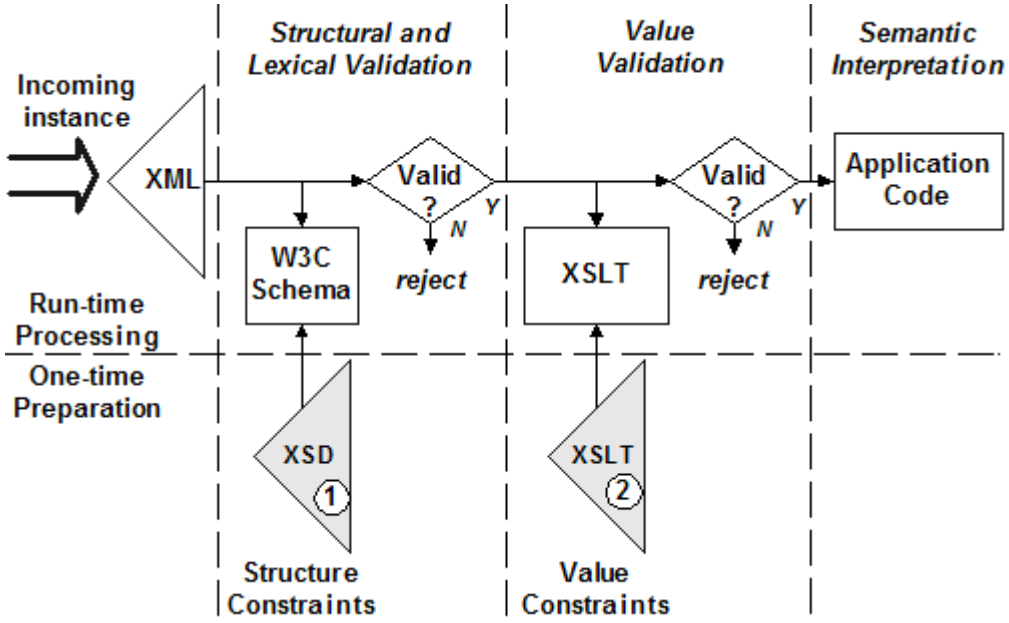


Figure 6, “Controlled vocabulary validation artifact creation” shows one way how the XSLT expression can be derived from an amalgam of the declarative resources published by BII: the document contexts of lists of code list and identifier information items and their associated value lists (labeled "3"), the value lists being referenced (labeled "4") and the business rules for BII document contents (labeled "5").

Figure 6. Controlled vocabulary validation artifact creation

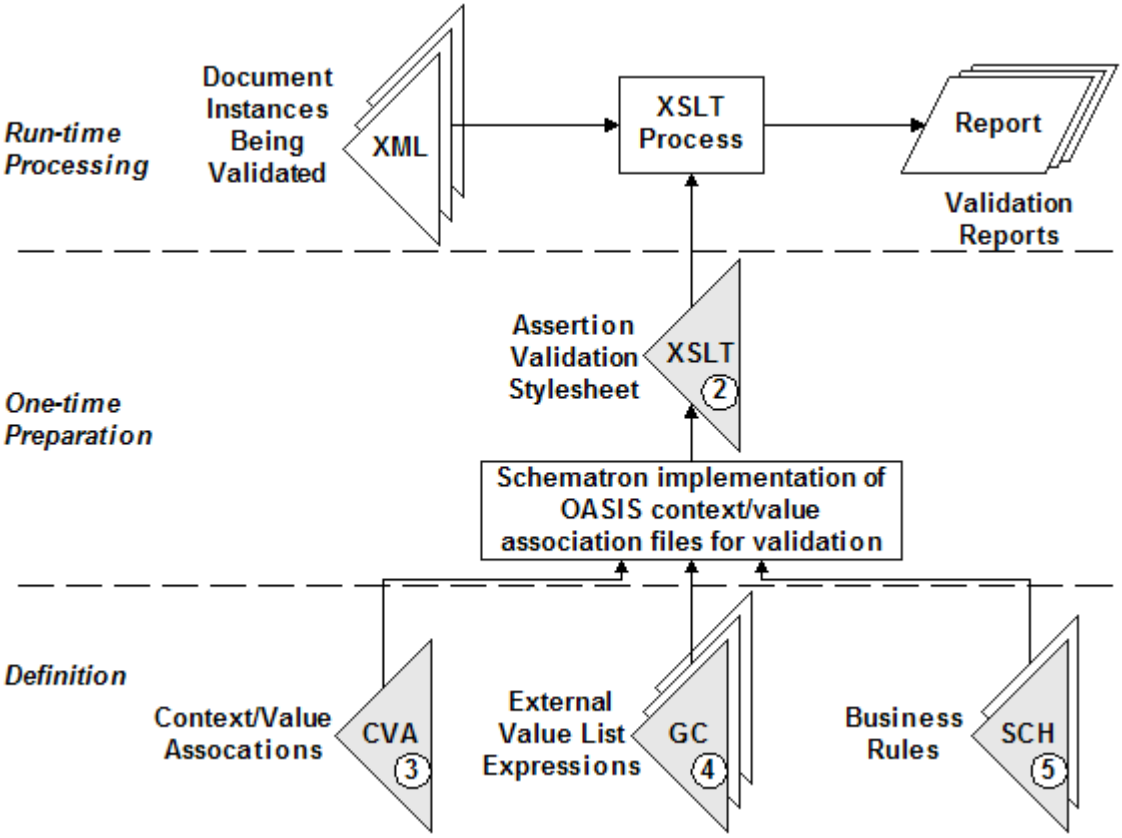
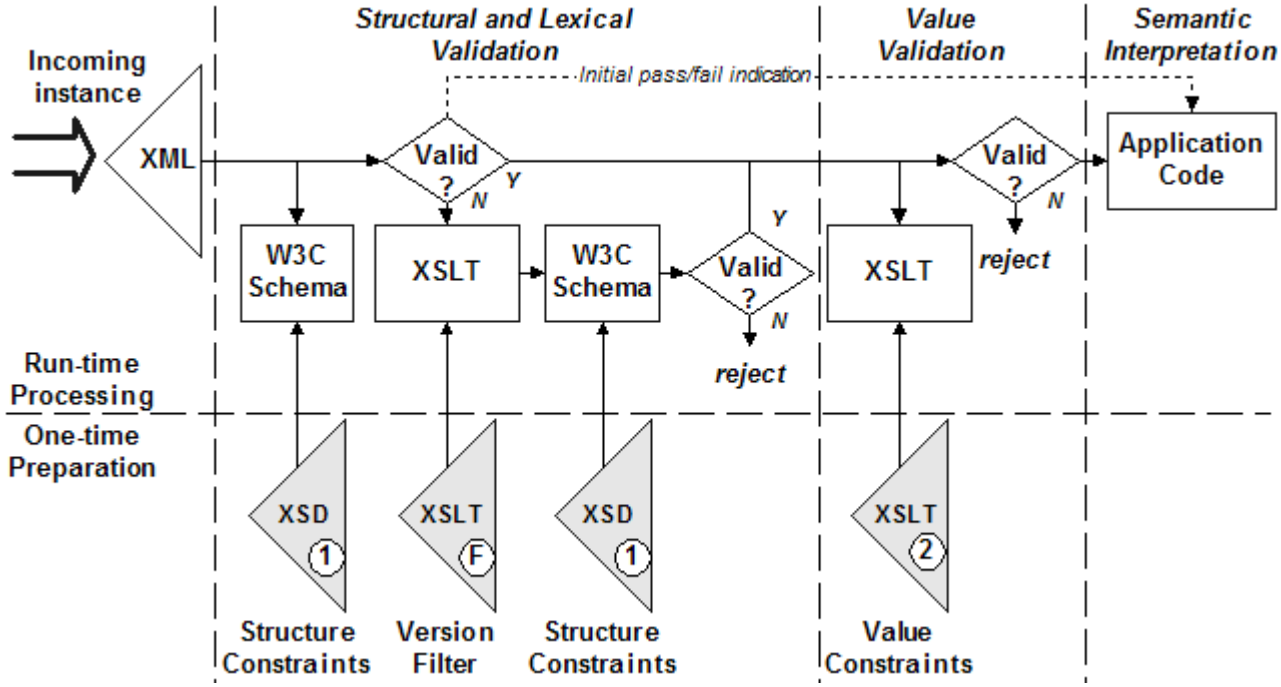


Figure 7, "Processing model accommodating versions of schemas" illustrates an alternative processing model for validating documents that accommodates forward compatibility with minor version releases of schemas. The version filter (labeled "F") is an example of a process that removes from a document every information item not recognized by the particular version of the schemas labeled "1". Users may choose to implement the filter using any process as this is not a declarative artefact but a procedural artefact based on the schema (or other expression of BII information entities) as the declarative artefact. Working group one could consider making such a procedural artefact available as a resource to implementers.

Figure 7. Processing model accommodating versions of schemas



8.1. Private code list and identifier values

Implementations and end users have the flexibility to extend or restrict the list of coded values allowed to be used in document contexts.

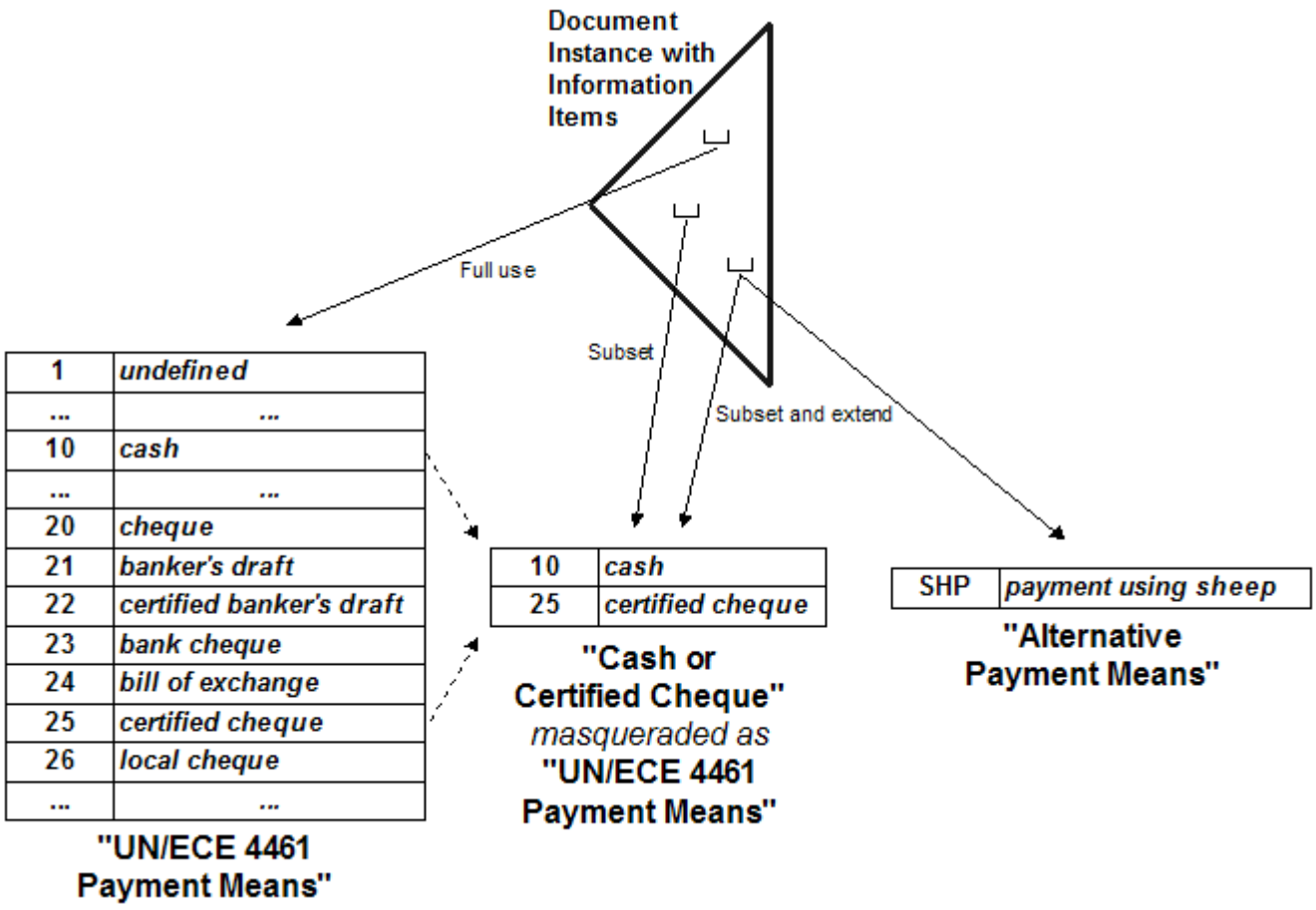
Extensions must be regarded as non-conformant changes to the document constraints. A trading partner using the published BII document constraints will be unfamiliar with any extensions used and will likely reject documents using them.

Restrictions can be used anywhere without impacting on the interoperability of a BII document. A trading partner utilizing the published BII document constraints will accept any document whose values are restricted to a subset of the published values.

Note that where BII does not constrain the values used in a code list or identifier value, the list of applicable values is effectively infinite. Any user imposing a restriction of the value of an item to a particular set of values will not violate the published BII document constraints for that item.

Figure 8, “Controlled vocabulary value applicability” illustrates an example of a user's choice to utilize a different set of values for three document contexts of payment means. In the left example the entire UN/ECE list of values is applicable. In the middle example the user has created their own list of only two standardized values, necessitating their own list-level meta data that is then masqueraded as the UN/ECE list for semantic identification. In the right-hand example the user allows values from both their own choice of standardized values as well as a non-standard extension of their own. A document using the extended value in the right-hand example could not be considered a BII document as an implementation of the published BII constraints would reject the use of the value.

Figure 8. Controlled vocabulary value applicability



Bibliography

[BII] Business Interoperability Initiative

[CLRTC] G. Ken Holman, Chair Code List Representation Technical Committee

[genericode] Tony Coates genericode, OASIS Code List Representation Technical Committee repository

[RELAX-NG] James Clark, Makoto Murata ISO/IEC 19757-2 RELAX-NG (Regular Language for XML)

[Schematron] Rick Jelliffe ISO/IEC 19757-3 Schematron, Document Schema Definition Languages (DSDL) Part 3, ISO/IEC JTC 1/SC 34/WG 1

[UBL 2.0] Jon Bosak, Tim McGrath, G. Ken Holman Universal Business Language (UBL) Version 2.0, OASIS UBL Technical Committee 2006

[W3C Schema] XML Schema Part 0: Primer, XML Schema Part 1: Structures, XML Schema Part 2: Datatypes 2004-10-28