Contents

Page

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work. In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 3.

ISO/IEC 19757-0 was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information Technology*, Subcommittee SC 34, Document Description and Processing Languages.

ISO/IEC 19757 consists of the following parts, under the general title *Document Schema Definition Languages (DSDL)*:

— *Part 0: Overview*

— *Part 1: Interoperability framework*

— *Part 2: Regular-grammar-based validation — RELAX NG*

— *Part 3: Rule-based validation — Schematron*

— *Part 4: Selection of validation candidates*

— *Part 5: Datatypes*

— *Part 6: Path-based integrity constraints*

— *Part 7: Character repertoire validation*

— *Part 8: Declarative document manipulation*

— *Part 9: Datatype- and namespace-aware DTDs*

Introduction

This International Standard defines the set of Document Schema Definition Languages (DSDL) used to specify one or more validation processes performed against Extensible Stylesheet Language (XML) documents. XML is an application profile of the Standard Generalized Markup Language (SGML) ISO 8879-1986.

A document model is an expression of the constraints dictated for the structure and content of documents to be validated with the model. A number of technologies have been developed through various formal and informal consortia since the development of Document Type Definitions (DTDs) as part of ISO 8879, notably by the World Wide Web Consortium (W3C) and the Organization for the Advancement of Structured Information Standards (OASIS). A number of validation technologies are standardized in DSDL to compliment those already available as standards or from industry.

To validate a structured document conforms to specified constraints in structure and content relieves the potentially many applications acting on the document from having to duplicate the task of confirming requirements have been met. Historically, such tasks and expressions have been developed and utilized in isolation without consideration for how the features and functionality available in other technologies can enhance the achievement of validation objectives.

Moreover, different design and use criteria may have led users to choose different validation technologies for separate portions of their information. Bringing together their information into an amalgam in a single XML document may prevent their existing document models from being used in isolation.

The main objective of this International Standard is to bring together varied validation-related tasks and expressions into a single extensible framework that allows each of the technologies implementing these individually to work in series and in parallel to produce a single or a set of validation results. The extensibility of DSDL accommodates validation technologies not yet designed or specified.

This multi-part standard integrates the best features of these proposals into a suite that:

— provides user control of names, order and repeatability of information objects (elements)

— allows users to identify restrictions on the co-concurrence of elements and element contents

— allows specific subsets of structured documents to be validated

— allows restrictions to be placed on the contents of specific elements, including restrictions based on the content of other elements in the same document

— allows the character set that can be used within specific elements to be managed, based on the application of the ISO 10646 Universal Multiple-Octet Coded Character Set (UCS)

— allows default values to be assigned to elements and attributes, and provides facilities for the incorporation of predefined fragments to be incorporated within documents

— allows SGML to be used to declare document structure constraints extends DTDs to include functions that are not currently provided for, such as namespaces and datatypes.

Document Schema Definition Languages (DSDL) — Part 0: Overview

1 Scope

This International Standard specifies new technologies and cites a number of existing technologies used in the process of validating the structure and content of XML documents. In addition, a framework is specified to choreograph the use of these validation technologies in order to orchestrate the production of a validation result.

DSDL defines the semantics, syntax and processing model for creating this validation result based on:

— A language specified in this International Standard for specifying the choreography of applying different validation technologies in the production of a validation result.

— A framework defined in this International Standard for orchestrating the application of different validation technologies to produce a validation result.

— Specifications of various validation technologies in different parts of this International Standard that can be used in isolation or within the DSDL framework.

— Citations to various validation technologies outside of this International Standard that can be used in isolation or within the DSDL framework.

This International Standard does not standardize a particular implementation of the framework defined or any validation technology specified in DSDL. DSDL expresses specifications to be performed by some processor that accepts an input document and produces a validation result.

Documents that are not in XML are not within the field of application of this International Standard.

All intermediate and final expressions of information when using DSDL shall be restricted to the XML Information Set and to XML documents. No expression of any concept supported by DSDL shall require anything beyond which can be expressed in an XML document.

This standard has the following parts, whose role is explained in the following sections:

— Part 1 — Interoperability Framework

— Part 2 — Regular-grammar-based Validation

— Part 3 — Rule-based Validation

— Part 4 — Selection of Validation Candidates

— Part 5 — Datatypes

— Part 6 — Path-based Integrity Constraints

— Part 7 — Character Repertoire Validation

— Part 8 — Declarative Document Manipulation

— Part 9 — Datatype and Namespace-aware DTDs

2 Normative references

The following normative documents contain provisions which, through reference in this text, constitute provisions of this part of ISO/IEC 19757. For dated references, subsequent amendments to, or revisions of, any of these publications

do not apply. However, parties to agreements based on this part of ISO/IEC 19757 are encouraged to investigate the possibility of applying the most recent editions of the normative documents indicated below. For undated references, the latest edition of the normative document referred to applies. Members of ISO and IEC maintain registers of currently valid International Standards.

IETF RFC 2396, *Uniform Resource Identifiers (URI): Generic Syntax*, Internet Standards Track Specification, August 1998, http://www.ietf.org/rfc/rfc2396.txt

SGML, *Standard Generalized Markup Language (SGML)*, ISO 8879-1986,

UCS, *Universal Multiple-Octet Coded Character Set (UCS)*, ISO 10646,

W3C XML, *Extensible Markup Language (XML) 1.0 (Second Edition)*, W3C Recommendation, 6 October 2000, http://www.w3.org/TR/2000/REC-xml-20001006

W3C XML-Infoset, *XML Information Set*, W3C Recommendation, 24 October 2001, http://www.w3.org/TR/2001/REC-xml-infoset-20011024/

W3C XML-Names, *Namespaces in XML*, W3C Recommendation, 14 January 1999, http://www.w3.org/TR/1999/REC-xml-names-19990114/

W3C XML Schema, *W3C XML Schema*, W3C Recommendation, 24 October 2001, http://www.w3.org/TR/2001/REC-xmlschema-0-20010502/

W3C XPath, *XML Path Language (XPath) Version 1.0*, W3C Recommendation, 16 November 1999, http://www.w3.org/TR/1999/REC-xpath-19991116

3 Terms and definitions

4 The role of document schemas

Document schemas provide machine-understandable models that can be used to validate the structure and contents of electronically marked-up documents. The Document Type Definition language defined within ISO 8879 provides facilities for:

— defining the names used to identify document elements

— identifying where document elements may appear in the document structure (model)

— identifying which elements were optional or repeatable (without limiting repeatability)

— identifying which markup tags are optional when then can be inferred through the model

— assigning properties (attributes) to document elements that can be used to control their processing, or can contain information that needs to be processed in conjunction with element contents

— defining default values for attributes

— defining and name repeatable segments of text (entities)

— identifying non-standard characters using user-assigned names or character numbers

— linking together different document structures defined in parallel sets of markup.

Document structures are defined in ISO 8879 in terms of "trees" of nested elements, though the standard also allows data sets to be defined as "graphs" of elements connected by means of unique identifiers and references to existing identifiers.

DTDs are not defined using the same components as documents, being defined in an efficient, sparse, notation. The notation can be preceded by a declaration of permitted character sets, which characters are assigned as control functions or otherwise ignored (shunned), which characters can be used as parts of names, or to identify the boundaries of markup, which strings used to be used to automatically identify markup points, and which optional functions are to be used within the DTD[1].

The W3C Extensible Markup Language (XML) uses an application profile of ISO 8879, known as WebSGML, and combines it with the ISO 10646 character set to produce a simple-to-implement, streamable, application of ISO 8879 for use over the Internet.

Various organizations have developed techniques to manage the structure of documents using XML markup. Some of these further subset the facilities provided in ISO 8879 to define document structures (e.g. ISO TR 22250, RELAX Core) but most of them also provide functions over and above those allowed within XML DTDs, including:

— control of the minimum and maximum number of times an element can occur at a particular point in the document structure (e.g. to control cardinality)

— restriction of the contents of particular elements or attributes to particular datatypes, patterns or internally defined lists of permitted elements

— provision for distinguishing the namespaces of element and attribute names so that DTD fragments can be incorporated into other DTDs without fear of name clashes

— identification of elements based on the path needed to reach them within the document structure

— validation of document structures by checking that elements conforming to particular paths exist

— provision of mechanisms for creating abstract stereotypes (similar to SGML architectural forms) that can be used to identify related classes of elements.

5 Other user requirements

The following additional functionality has been identified as being required by users:

— The ability to control the character set permitted within the contents of a particular type of element or attribute, or within specified sets of elements within the document model.

— The ability to restrict the range of entries conforming to a particular use of a datatype within a specific element or attribute.

— The ability to restrict element or attribute contents to values specified in either internally defined or externally defined lists of permitted values.

— The ability to restrict the set of permitted values in one element or attribute based on the contents of another element or attribute (e.g. not Sex=Male and Condition=Pregnant).

— The ability to generate compact forms of schemas that are easily readable by humans, and to use such compact representations to generate schemas or DTDs that can be used to validate documents.

— The ability to visualize schemas using navigable diagrammatic representations.

Question: What other entries need to be added to this list?

[1] The character definition rules predate the development of the ISO 10646 Universal Multiple-Octet Coded Character Set and are to some extend made redundant by this standard.

6 Choreography

The various parts of the DSDL standard are designed to be choreographed to satisfy a declarative expression of the validation requirements, without the need to use processes in a predefined sequence. Some parts of the standard will, however, need to be applied before others. For example, validation of the contents of a specific element will require prior identification of element boundaries and nesting, while the validation of the relationship between elements may require that the document structure be validated first so that the paths specified can be checked accordingly.

Consider the example in Figure 1 where two different results can be determined from two different applications of technology to the validation process: validating after or before processing XInclude. The order of these two steps may be critical in the correct processing of the information in the instance.

In a more complex example, consider Figure 2 where two different technologies must be applied to separate portions of the one document. In this case, one part of the input document must be validated by a W3C XML Schema while the other part of the input document must be validated by a RELAX-NG schema. The validation result, orchestrated by the DSDL framework, would express the consolidated validation of all steps.

7 Path-based addressing

The non-hierarchical links between information items in a structured resource can be identified by addressing the items and expressing the relationship between them found in the document tree. The addressing mechanism includes hierarchy-based paths of steps along the tree's branches to the information item being addressed.

Paths are based on:

— a method of identifying information items based on:

  — the ancestry of the information item

  — other mechanisms not based on the tree (e.g. unique identifier values)

  — an extensible basis for supporting mechanisms not currently available

— a method of describing relationships that are not hierarchical

A number of Parts utilize this concept of a path to address components of document instances.

Paths are used in both parts 3 and 6.

Should this be a separate part referenced by the others? Would it belong in Part 0 with scope over all the others?

8 Overview of all of the parts

8.1 The Interoperability Framework

The DSDL Interoperability Framework is a language for choreographing the validation and pre-validation transformation processes described in ISO/IEC 19757-2 to ISO/IEC 19757-9.

Within this framework:

— pre-validation transformations may be used to isolate and normalize documents before validation

— multiple validations and transformations may be applied to the same document

— transformations may split a document into multiple resulting documents

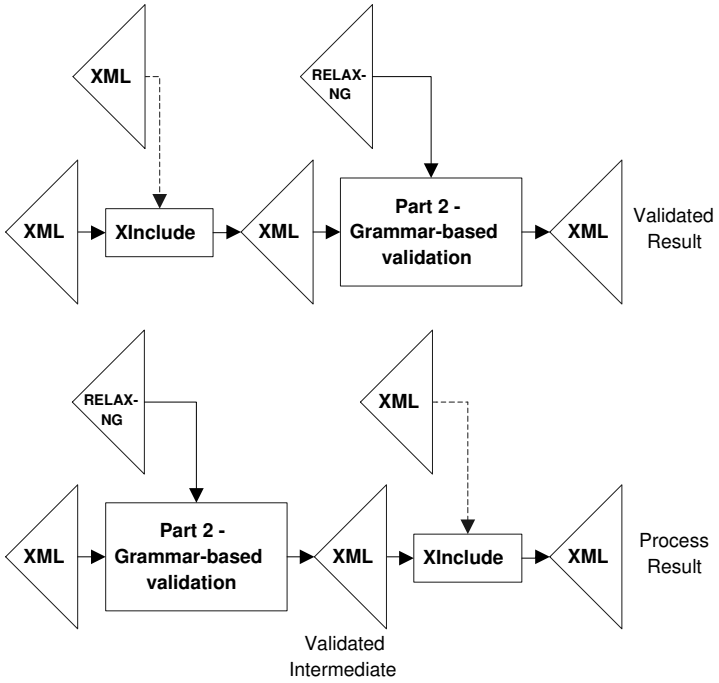

Figure 1: Two different orders of application of technology

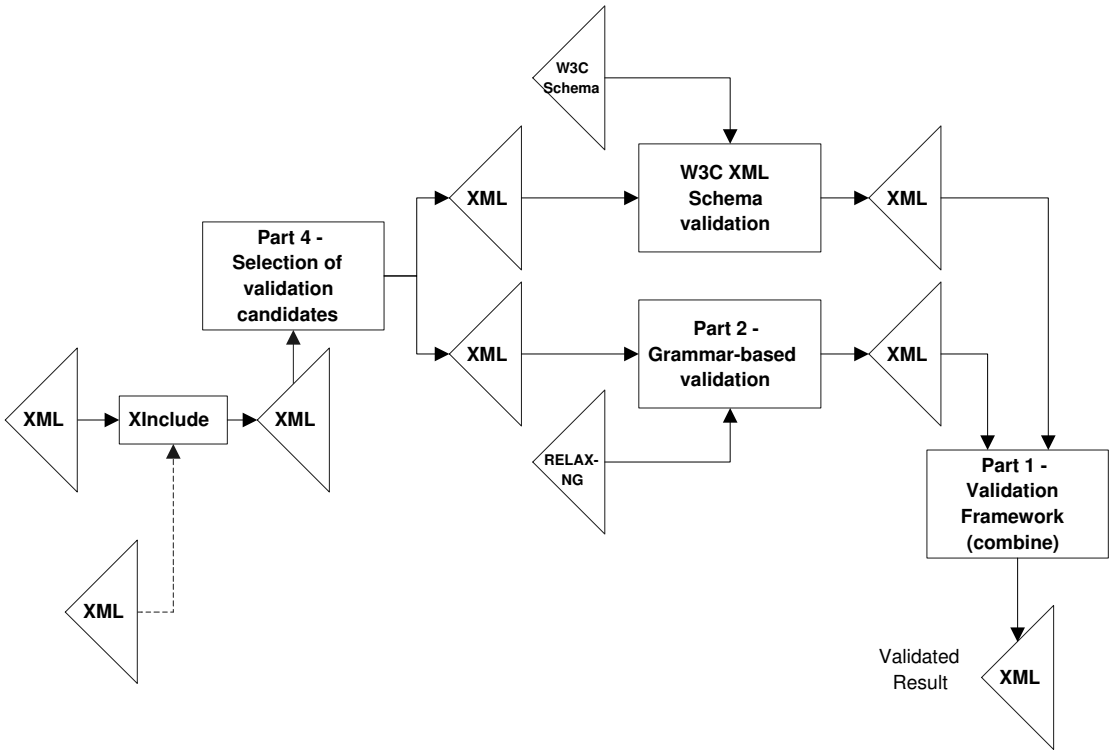

Figure 2: A multi-step validation process

— users can customize the validation reports which can be output as XML and further processed by other applications

— a mechanism is provided to call specific parsers which read non XML sources (and XML sources that can't be identified by a single URI) into XML infosets. Examples of such sources include SGML and HTML documents, RDBMS query results, CSV documents and Web Services query results

Note that not only DSDL-specified processes can be identified within the framework. This part illustrates how other technologies, in particular W3C XML Schema and XSLT, can be utilized from within the framework.

8.2 Regular-grammar-based Validation

Regular-grammar-oriented schema languages validate that the structure and content of information items in a document instance conforms to a model described by a tree grammar.

This Part includes a syntax for specifying:

— the elements that can make up the hierarchy

— the attributes that can be assigned to elements

— the identity of datatypes, their permitted ranges and the permitted values

Tree grammars are characterized by the specification of patterns, and validation is based on the replacement of patterns by their definitions being matched in the stream being analyzed.

Note that not only DSDL-specified datatypes can be identified when using the schema language defined in this part. This part illustrates how other datatype definitions, in particular W3C XML Schema Data types, can be utilized as part of the validation process.

This Part is based on the OASIS RELAX NG specification. Other grammar-oriented schema languages may be defined in the evolution of DSDL as separate parts of this International Standard.

8.3 Rule-based Validation

This Part provides a syntax for specifying:

— the relationship between contents in different parts of document instances identified through the use of path expressions

This Part is based on the Schematron specification. Other Rule-based validation languages may be defined in the evolution of DSDL as separate parts of this International Standard.

8.4 Selection of validation candidates

This part provides an XML-based language for selecting specific elements within a document instance that are to be validated independently. Such elements are called validation candidates.

Selection methods include:

— namespace-based selection, which is controlled by conditions on namespaces of elements

— attribute-based selection, which is controlled by conditions on attributes of attributes

Schema languages other than DSDL (for example RDF Schema[1] and the Topic Map Constraint Language[2]) may be used for validating selected validation candidates. For example, an XHTML document containing metadata (RDF or topic map) can be decomposed into an XHTML validation candidate and metadata validation candidate and validated independently.

It is outside the scope of this part to specify which schema and schema language is used for validating validation candidates.

The XML-based language of this part may be used to create an independent XML document or may be used to create a part of an XML document. Specifically, when a DSDL framework is represented by an XML document, it may reference to or contain descriptions in this XML-based language.

8.5 Datatypes

This Part defines:

— a set of standardized named datatypes (e.g. `integer`)

— a set of parameters and their values for each datatype (e.g. minimum and maximum values)

— a set of constraints describing a possibly infinite set of strings representing values of the data type

This Part has been developed from the set of primitive datatypes and their facets defined in Part 2 of the W3C XML Schema specification.

8.6 Path-based integrity constraints

This Part defines:

— the expression of identity and integrity constraints on components of document instances identified through the use of path expressions

8.7 Character Repertoire Validation

This Part provides a syntax for:

— defining named subsets of the ISO 10646 character set

— identifying which named character set shall be used to validate the content of a specific element or attribute

8.8 Declarative Document Manipulation

This Part provides a syntax for:

— assigning a default value to the contents of a specific type of element or attribute

— defining named entities of predefined data elements that can be used to include template data within a document instance

— renaming elements and attributes in specific locations within the document model

Under consideration for this part are Architectural Form and Architecture Support Attribute approaches.

8.9 Namespace and datatype aware DTDs

This part specifies how Document Type Definition (DTD) syntax can be interpreted to validate documents that make full use of XML Namespaces and Part 5 Datatypes. This will help to preserve the investment that individuals and organizations have made in DTD development and deployment. It will also help those converting between DTD expressions and other schema languages.

The specification will not require documents using the schema language to violate XML's well-formedness or validity checks. Part 9 will be just another schema language that may be applied alongside others.

Bibliography

[1] *RDF Vocabulary Description Language 1.0: RDF Schema*, http://www.w3.org/TR/rdf-schema

[2] *Topic Map Constraint Language*, http://www.w3.org/TR/rdf-schema

Summary of editorial comments:

[5] Other user requirements

Question: What other entries need to be added to this list?

[7] Path-based addressing

Paths are used in both parts 3 and 6.

Should this be a separate part referenced by the others? Would it belong in Part 0 with scope over all the others?