

**Contents**

Page

<b>Foreword</b> .....	<b>iv</b>
<b>Introduction</b> .....	<b>v</b>
<b>1 Scope</b> .....	<b>1</b>
<b>2 Normative references</b> .....	<b>1</b>
<b>3 Terms and definitions</b> .....	<b>1</b>
<b>4 Schematron Query Language Binding</b> .....	<b>1</b>
<b>5 Conformance</b> .....	<b>3</b>
<b>5.1 Simple Conformance</b> .....	<b>3</b>
<b>5.2 Full Conformance</b> .....	<b>3</b>
<b>Annex A (informative) Use Cases</b> .....	<b>4</b>
<b>Annex B (informative) Example Schemas for Use-Cases</b> .....	<b>6</b>

## 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 19757-7 was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information Technology*, Subcommittee SC 34, Document Description and Processing Languages.

- *Part 0: Overview*
- *Part 1: Interoperability framework*
- *Part 2: 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

The structure of this standard is as follows. Clause 4 specifies the schema language as a particular query language binding of Part 3 (Schematron. ) Clause 5 describes conformance requirements for implementations of character repertoire validators. Finally, non-normative annexes provide motivating use-cases and examples.



# Document Schema Definition Languages (DSDL) — Part 7: Character repertoire validation

## 1 Scope

This standard specifies a schema language for declaring and validating the allowed character repertoire in data content, attribute values, identifiers and other markup in XML documents. The language is specified as a particular query language binding of Part 3 (Schematron.)

The schema language use XSLT path expressions to declare contexts for assertions, and XSD character classes to specify repertoires.

This standard establishes requirements for schemas and specifies when an XML document matches the patterns specified by the schema.

## 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.

The following referenced documents are indispensable for the application of this standard. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

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

XSLT, *XSL Transformations (XSLT) Version 1.0*, W3C Recommendation, <http://www.w3.org/TR/1999/REC-xslt-19991116>

XSD, *XML Schema Part 2: Datatypes*, W3C Recommendation, <http://www.w3.org/TR/2001/REC-xmlschema-2-20010502/#nt-charGroup>

## 3 Terms and definitions

The definitions of Part 1, Part 2 and Part 3 also apply to this standard.

### character class

A grouping of characters, especially into named groups according to some property of the characters.

### character repertoire

The characters which may validly be used in some text node.

### regular expression

An artificial language or group of dialects for expressing patterns in sequences of characters.

## 4 Schematron Query Language Binding

A schema conforming to this standard shall be a correct Schematron schema according to Part 3 (Schematron.)

The value of the `language` attribute of the Schematron schema element shall be `xpath-charrep`, in any mix of upper and lower case letters.

NOTE This standard reserves following query language names without further definition. Implementations which use different query language bindings shall use one of these names if appropriate: `stx-charrep`, `xslt1.1-charrep`, `exslt-charrep`, `xslt2-charrep`, `xpath-charrep`, `xpath2-charrep`, `xquery-charrep`.

The following binding shall be used:

- The rule context shall be interpreted according to the production 1 of XSLT, as returning any kind of XPath node when applied using the semantics of Part 3 (Schematron).
- The assertion test shall be interpreted according to production 13 of XSD, returning a boolean. Each string shall be tested on a character-by-character basis. Newlines and tabs shall be removed from the query before it is used to allow longer strings; consequently these characters must be entered using the delimited forms `\n``\t``\r` or the character classes. It shall not be an error if the same character or class is entered more than once.
- The name query shall be interpreted according to production 14 of XPath, as returning a string value.
- The value-of query shall be interpreted according to production 14 of XPath, as returning a string value.
- The `let` element shall not be used.
- Abstract patterns shall not be used.

NOTE The nodes allowed as subjects for assertions in this standard could be different from the nodes allowed by the default query language binding of Part 3 (Schematron).

The XPath data model shall be used. For the purpose of this standard an XPath string is an XSD string. For each node type, the strings to be tested are as follows:

- For an element node: the string-values of each text node of that element node, but not the data content of any descendents.
- For an attribute node: the string-value of the attribute, as would be accessible through the XPath `text()` function for that subject.
- For an comment node: the string-value of the comment, as would be accessible through the XPath `text()` function for that subject.
- For an processing instruction node: the string-value of the processing instruction, as would be accessible through the XPath `text()` function for that subject.
- For name nodes: the name as a string.
- For any other nodes, including the document root node: error.

The XSLT `key` element shall not be used.

The order in which nodes and characters are validated is not specified by this standard.

NOTE Parallel constraints may be expressed using separate assertions.

## 5 Conformance

### 5.1 Simple Conformance

A simple-conformance implementation has the same requirements as a simple-conformance implementation of Part 3 (Schematron.)

### 5.2 Full Conformance

A full-conformance implementation has the same constraints as a full-conformance implementation of Part 3 (Schematron) and the following additional requirements.

- The schema has a language binding attribute with a value terminated by the string of `-charrep`.
- The schema with a language binding attribute with value `xslt-charrep` conforms to the language binding in this standard.

## Annex A (informative)

### Use Cases

Motivating use-cases for the schema language include:

- restricting the generic identifiers of elements to ASCII characters;
- ensuring that a Dutch document contains characters only used in typical Dutch documents; the constraint applies to mixed content and element content;
- checking that a document does not use any Latin combining characters;
- declaring that data content in a Japanese document shall not contain *half-width katakana* characters;
- providing information to alert publishing staff if the data content of a document contains characters outside the Unicode Basic Multilingual Plane, surrogate characters, or Private Use Area characters;
- verifying that the data content in a scientific document uses the Unicode character for micro symbol not the Greek small letter mu;
- verifying a school text book that data content of Japanese *ruby* annotations does not contain Han ideographs and that other data content of elements should contain only the restricted repertoire used for schools; because the repertoire is large, it must be declared in an external library and referenced; and
- verifying that an attribute value giving a person's name in a Chinese document only uses approved characters.

Motivating use-cases for the schema language do not include:

- constraints on parts of a string, such as that an attribute should start with a certain character;
- semantic constraints requiring analysis of the particular string, such as that that an attribute may contain letters or numbers but not both;
- repertoire constraints between different strings in the document, such as that an element can only use the character repertoire as used in some other part of the document; and
- constraints involving arithmetic operations, such as that the sum of all code values in the string should not exceed 300.

As well, certain kinds of constraints are out-of-scope for this standard:

- the maximum and minimum length of strings;
- the character encoding (character set) used for an entity;
- the use of standard entities, numeric character references or literal characters;
- that the characters of a Thai document are ordered correctly; and
- that the initial characters of a portion of a string marked up with an entity or included by some macro mechanism is not a combining character.



## Annex B (informative)

### Example Schemas for Use-Cases

Restricting the generic identifiers of elements to ASCII characters

```
<sch:rule context="*/@name(">
  <sch:assert test="[\p{IsBasicLatin}]">
    Generic identifiers of elements should be ASCII repertoire.
  </sch:assert>
</sch:rule>
```

Ensuring that a Dutch document contains characters only used in typical Dutch documents; the constraint applies to mixed content and element content

```
<sch:rule context="*[/*[@xml:lang='nl']]">
  <sch:assert test="\p{IsBasicLatin}\p{IsLatin-1Supplement}
    &#x132;&#x133;\p{IsGeneralPunctuation}\p{IsCurrencySymbols}">
    If this document is a Dutch document, it should have only characters
    used in typical Dutch publishing.
  </sch:assert>
</sch:rule>
```

Checking that a document does not use any Latin combining characters

```
<sch:rule context="*">
  <sch:assert test="^\p{Lm}">
    This document should not use any Latin combining characters.
  </sch:assert>
</sch:rule>
```

Declaring that data content in a Japanese document shall not contain *half-width katakana* characters

```
<sch:rule context="* | @">
  <sch:assert test="^\p{IsSmallFormVariants}">
    Elements and attributes should not contain half-width katakana characters.
  </sch:assert>
</sch:rule>
```

Verifying a school text book that data content of Japanese *ruby* annotations does not contain Han ideographs

```
<sch:rule context="rb">
  <sch:assert test="^\p{IsCJKUnifiedIdeographs}">
    Ruby annotations should not contain Han ideographs.
  </sch:assert>
</sch:rule>
</sch:rule>
```

Providing information to alert publishing staff if the data content of a document contains characters outside the Unicode Basic Multilingual Plane, surrogate characters, or Private Use Area characters

```

<sch:rule context="*">
  <sch:assert test="&#x01;-&#xFFEF;">This document should not
    contain characters outside the Unicode Basic Multilingual Plane.
  </sch:assert>
  <sch:assert test="^&#xD800;-&#xDFFF;">This document should not
    contain characters surrogate characters.
  </sch:assert>
  <sch:assert test="^\p{Co}">This document should not
    contain Private Use Area characters.
  </sch:assert>
</sch:rule>

```

Verifying that the data content in a scientific document uses the Unicode character for micro symbol not the Greek small letter mu

```

<sch:rule context="*">
  <sch:assert test="^&#x3BC;">
    The micro symbol should be used, not the Greek small letter mu.
  </sch:assert>
</sch:rule>

```

Verifying that other data content of elements should contain only the restricted repertoire used for schools; because the repertoire is large, it must be declared in an external library and referenced

TO DO: define an abstract rule. Include the rule using some inclusion mechanism.

The preceding schema fragments should be placed in the following wrapper for them to be correct schemas.

```

<sch:schema xmlns:sch="http://www.ascc.net/xml/schematron" language="xslt-charrep" >
  <sch:title>Examples of Use Cases</sch:title>
  <sch:pattern name="Example">
    ...
  </sch:pattern>
</sch:schema>

```

Summary of editorial comments: