ISO/IEC JTC 1/SC 34

Date: 2005-2-18

ISO/IEC CD 19757-7

ISO/IEC JTC 1/SC 34/WG 1

Secretariat: Standards Council of Canada

Document Schema Definition Languages (DSDL) — Part 7: Character Repertoire Validation Language

Warning

This document is not an ISO International Standard. It is distributed for review and comment. It is subject to change without notice and may not be referred to as an International Standard.

Recipients of this document are invited to submit, with their comments, notification of any relevant patent rights of which they are aware and to provide supporting documentation.

Document type: International Standard
Document subtype:
Document stage: (30) Committee
Document language: E

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 2.

ISO/IEC 19757-7 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 1: Overview*

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

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

— *Part 4: Namespace-based validation dispatching language — NVDL*

— *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*

— *Part 10: Validation management*

Introduction

This International Standard defines a set of Document Schema Definition Languages (DSDL) that can be used to specify one or more validation processes performed against Extensible Markup Language (XML) documents. A number of validation technologies are standardized in DSDL to complement those already available as standards or from industry.

The main objective of this International Standard is to bring together different validation-related technologies to form a single extensible framework that allows technologies to work in series or 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 part of ISO/IEC 19757 provides a schema language for describing collections of ISO/IEC 10646 characters. Schemas in this language may be referenced from other schema languages.

The structure of this part of ISO/IEC 19757 is as follows. Clause 5 introduces kernels and hulls of character collections. Clause 6 describes the syntax of CRVL schemas. Clause 7 describes the semantics of a correct CRVL schema; the semantics specify when a character is contained by a collection described by a CRVL schema.Clause 8 describes conformance requirements for CRVL validators.

Document Schema Definition Languages (DSDL) — Part 7: Character Repertoire Validation Language

1 Scope

This part of the International Standard specifies a Character Repertoire Validation Language (CRVL). A CRVL schema describes a collection of ISO/IEC 10646 characters.

2 Normative references

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

Each of the following documents has a unique identifier that is used to cite the document in the text. The unique identifier consists of the part of the reference up to the first comma.

RELAX NG, *ISO/IEC 19757-2, Document Schema Definition Languages (DSDL) — Part 2: Grammar-based validation — RELAX NG*

W3C XML, *Extensible Markup Language (XML) 1.0 (Third Edition)*, W3C Recommendation, 04 February 2004, available at <http://www.w3.org/TR/2004/REC-xml-20040204/>

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

W3C XML Schema Part 2, *XML Schema Part 2: Datatypes*, W3C Recommendation, 02 May 2001, available at <http://www.w3.org/TR/2001/REC-xmlschema-2-20010502/>

IETF RFC 3986, *Uniform Resource Identifiers (URI): Generic Syntax*, Internet Standards Track Specification, January 2005, available at <http://www.ietf.org/rfc/rfc3987.txt>

IETF RFC 3987, *Internationalized Resource Identifiers (IRIs)*, Internet Standards Track Specification, January 2005, available at <http://www.ietf.org/rfc/rfc3987.txt>

ISO/IEC 10646, *Universal multiple-octet coded Character Set*

Unicode, *The Unicode standard*

3 Terms and definitions

For the purposes of this part of ISO/IEC 19757, the following terms and definitions apply.

3.1
character collection
a set of characters

4 Notation

— in(*x*, *A*) character *x* is in collection *A*

- — notin(*x*, *A*) character x is not in collection *A*
- — in(*x*, *A*) it is unknown whether character *x* is in collection *A* or not

5 Kernel and hull

A kernel contains characters that are guaranteed to be in the collection; the collection may contain other characters. A hull gives an outer boundary so that characters which are not in the hull are guaranteed not to be in the collection; some characters in the hull may not actually be in the collection.

NOTE See 2.2 in A Notation for Character Collections for the WWW[1] .

6 Syntax

An CRVL schema in the full syntax shall be an XML document valid against the following RELAX NG schema in the compact syntax.

```
default namespace = "toBeSupplied"

start = element characterCollection {
          attribute minUcsVersion {text}?,
          attribute maxUcsVersion {text}?,
          coll}

coll = hull | kernel | union | intersection | difference |
      alt | ref | namedCollection | charGroup

hull = element hull { coll }
kernel = element kernel { coll }

union = element union { coll* }
intersection = element intersection { coll* }
difference = element difference { coll, coll }
alt = element alt { coll* }

ref = element ref {attribute href { xsd:anyURI }}
namedCollection = element namedCollection { attribute ns { xsd:anyURI } }

charGroup = element charGroup { text }
```

The content of an charGroup element matches charGroup as specified in W3C XML Schema Part 2.

NOTE The following rules are copied from . by the following rules.

The semantics of [29] thru [37] depend on the version of the Unicode standard.

```
[12] charClassExpr ::= '[' charGroup ']'
[13] charGroup ::= posCharGroup | negCharGroup | charClassSub
[14] posCharGroup ::= ( charRange | charClassEsc )+
[15] negCharGroup ::= '^' posCharGroup
[16] charClassSub ::= ( posCharGroup | negCharGroup )
                   '-' charClassExpr
[17] charRange ::= seRange | XmlCharIncDash
[18] seRange ::= charOrEsc '-' charOrEsc
[20] charOrEsc ::= XmlChar | SingleCharEsc
[21] XmlChar ::= [^\#x2D#x5B#x5D]
[22] XmlCharIncDash ::= [^\#x5B#x5D]
[23] charClassEsc ::= ( SingleCharEsc | MultiCharEsc
                    | catEsc | complEsc )
[24] SingleCharEsc ::= '\' [nrt\|.?*+(){}#x2D#x5B#x5D#x5E]
[25] catEsc ::= '\p{' charProp '}'
[26] complEsc ::= '\P{' charProp '}'
[27] charProp ::= IsCategory | IsBlock
[28] IsCategory ::= Letters | Marks | Numbers
                  | Punctuation | Separators | Symbols | Others
[29] Letters ::= 'L' [ultmo]?
[30] Marks ::= 'M' [nce]?
[31] Numbers ::= 'N' [dlo]?
[32] Punctuation ::= 'P' [cdseifo]?
[33] Separators ::= 'Z' [slp]?
[34] Symbols ::= 'S' [mcko]?
[35] Others ::= 'C' [cfon]?
[36] IsBlock ::= 'Is' [a-zA-Z0-9#x2D]+
[37] MultiCharEsc ::= '\' [sSiIcCdDwW]
```

7 Semantics

7.1 General

<characterCollection minUcsVersion="*min*" maxUcsVersion="*max*"/> *A* </characterCollection> is meant to represent a character collection *A* where the Unicode version greater than *min* (if any) and less than *max* (if any) shall be used.

Given a character x and a character collection *A*, either in(*x*, *A*), notin(*x*, *A*), or unknown(*x*, *A*) holds.

7.2 <hull>*A*</hull>

— in(*x*,<hull>*A*</hull>) does not hold.

— notin(*x*,<hull>*A*</hull>) when notin(*x*, *A*).

— unknown(*x*,<kernel>*A*</kernel>) when in(*x*, *A*) or unknown(*x*, *A*).

NOTE The hull of *A* is used as that of <hull>*A*</hull>. The kernel of <hull>*A*</hull> is empty.

7.3 <kernel>*A*</kernel>

— in(*x*,<kernel>*A*</kernel>) when in(*x*, *A*).

— notin(*x*,<kernel>*A*</kernel>) does not hold.

— unknown(*x*,<kernel>*A*</kernel>) when notin(*x*, *A*) or unknown(*x*, *A*).

NOTE The kernel of *A* is used as that of <kernel>*A*</kernel>. The hull of <kernel>*A*</kernel> is universal.

7.4 <union>*A B* ...</union>

— in(*x*, <union>*A B*</union>) when in(*x*, *A*) or in(*x*, *B*).

- notin(*x*, <union>*A B*</union>) when notin(*x*, *A*) and notin(*x*, *B*).

- unknown(*x*, <union>*A B*</union>) otherwise.

7.5 <intersection>*A B* ...</intersection>

- in(*x*, <intersection>*A B*</intersection>) when in(*x*, *A*) and in(*x*, *B*).

- notin(*x*, <intersection>*A B*</intersection>) when notin(*x*, *A*) or notin(*x*, *B*)

- unknown(*x*, <intersection>*A B*</intersection>) otherwise.

7.6 <difference> *A B* </difference>

- in(*x*, <difference>*A B*</difference>) when in(*x*, *A*) and notin(*x*, *B*)

- notin(*x*, <difference>*A B*</difference>) when notin(*x*, *A*) or in(*x*, *B*)

- unknown(*x*, <difference>*A B*</difference>) otherwise.

7.7 <alt>*A B*</alt>

- in(*x*, <alt>*A B*</alt>) when in(*x*, *A*) or in(*x*, *B*)

- notin(*x*, <alt>*A B*</alt>) when notin(*x*, *A*) or notin(*x*, *B*)

- unknown(*x*, <alt>*A B*</alt>) when unknown(*x*, *A*) and unknown(*x*, *B*)

NOTE See the semantics of ALT in A Notation for Character Collections for the WWW[1] .

7.8 <ref href="*uri*"/>

A CRVL schema (say A) shall be obtained by deferencing uri.

- in(*x*,<ref href="*uri*"/>) when in(*x*, *A*).

- notin(*x*,<ref href="*uri*"/>) when notin(*x*, *A*).

- unknown(*x*,<ref href="*uri*"/>) when unknown(*x*, *A*).

7.9 <namedCollection name="*uri*"/>

Some 3.1 ("Well-known collections") in A Notation for Character Collections for the WWW[1] .

7.10 <charGroup>*charGroup*</charGroup>

— in(*x*, <charGroup>*charGroup*</charGroup>) when x matches *charGroup*.

— notin(*x*, <charGroup>*charGroup*</charGroup>) when *x* does not match *charGroup*.

— unknown(*x*, <charGroup>*charGroup*</charGroup>) does not hold.

8 Conformance

To be supplied.

Bibliography

[1] *A Notation for Character Collections for the WWW*, W3C Note, 14 January 2000, available at <http://www.w3.org/TR/charcol>