FINAL
DRAFT

INTERNATIONAL
STANDARD

ISO/IEC FDIS 19757-7

ISO/IEC JTC **1**/SC **34**

Secretariat: JISC

Voting begins on:
2020-05-12

Voting terminates on:
2020-07-07

Information technology — Document Schema Definition Languages (DSDL) —

Part 7: Character Repertoire Description Language (CREPDL)

Technologies de l'information — Langages de définition de schéma de documents (DSDL) —

Partie 7: Langage de description de répertoire de caractères (CREPDL)

Reference number
ISO/IEC FDIS 19757-7:2020(E)

COPYRIGHT PROTECTED DOCUMENT

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Fax: +41 22 749 09 47
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

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.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents) or the IEC list of patent declarations received (see http://patents.iec.ch).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

This document was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 34, *Document description and processing languages*.

This second edition cancels and replaces the first edition (ISO/IEC 19757-7:2009), which has been technically revised. It also incorporates the Technical Corrigendum ISO/IEC 19757-7:2009/Cor 1:2015.

The main changes compared to the previous edition are as follows:

— addition of validation of grapheme clusters such as 'n' followed by COMBINING GRAVE ACCENT (U+0300) and a CJK unified ideograph followed by a variation selector.

— addition of the Unicode Ideographic Variation Database as a registry.

A list of all parts in the ISO/IEC 19757 series can be found on the ISO website.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

ISO/IEC 19757 (all parts) 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 ISO/IEC 19757 (all parts) 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 document provides a language for describing character repertoires. Descriptions in this language can be referenced from schemas. Furthermore, they can also be referenced from forms and stylesheets.

Descriptions of character repertoires doesn't need to be exact. Non-exact descriptions are made possible by kernels and hulls, which provide the lower and upper limits, respectively.

The structure of this document is as follows. Clause 5 provides an informal overview of CREPDL. Clause 6 specifies the syntax of CREPDL schemas. Clause 7 specifies the semantics of a correct CREPDL schema; the semantics specify when a code point or code point sequence is in a character repertoire described by a CREPDL schema. Clause 8 defines the behaviour of CREPDL processors. Finally, Annex A describes differences of conformant CREPDL processors; Annex B provides examples of CREPDL schemas.

Although the first edition was restricted to the validation of characters, this edition can also enable the validation of grapheme clusters such as 'n' followed by COMBINING GRAVE ACCENT (U+0300) and a CJK unified ideograph followed by a variation selector.

CREPDL schemas conformant to the first edition do not conform to this edition. In particular, this edition changes the namespace name for CREPDL schemas.

Information technology — Document Schema Definition Languages (DSDL) —

Part 7:
Character Repertoire Description Language (CREPDL)

1 Scope

This document specifies a Character Repertoire Description Language (CREPDL). A CREPDL schema describes a character repertoire. A stream of UCS code points can be validated against a CREPDL schema.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements 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.

ISO/IEC 10646, *Information technology — Universal Multiple-Octet Coded Character Set (UCS)*

ISO/IEC 19757-2, *Information technology — Document Schema Definition Language (DSDL) — Part 2: Regular-grammar-based validation — RELAX NG*

ISO/IEC 19757-4, *Information technology — Document Schema Definition Languages (DSDL) — Part 4: Namespace-based Validation Dispatching Language (NVDL)*

W3C XML, *Extensible Markup Language (XML) 1.0 (Fourth Edition)*, W3C Recommendation, 16 August 2006, available at http://www.w3.org/TR/2006/REC-xml-20060816

W3C XML-Names, *Namespaces in XML (Second Edition)*, W3C Recommendation, 16 August 2006, available at http://www.w3.org/TR/2006/REC-xml-names-20060816

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

Charsets I.A.N.A. *IANA CHARACTER SETS*, available at http://www.iana.org/assignments/character-sets

Unicode, *The Unicode Standard*, The Unicode Consortium, available at http://www.unicode.org/

CLDR, *Unicode Common Locale Data Repository*, The Unicode Consortium, available at http://www.unicode.org/cldr/

UAX29, *Unicode Standard Annex #29: Unicode Text Segmentation*, The Unicode Consortium, available at http:// unicode.org/reports/tr29/

UTS35, *Unicode Technical Standard #35: Unicode Locale Data Markup Language (LDML)*, The Unicode Consortium, available at https://www.unicode.org/reports/tr35/

UTS37, *Unicode Technical Standard #37: Unicode Ideographic Variation Database*, The Unicode Consortium, available at http://www.unicode.org/reports/tr37/

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at http://www.electropedia.org/

3.1
CREPDL processor
computer program that validates a stream of code points not containing high- or low-surrogate code points against *CREPDL schemas* (3.2)

3.2
CREPDL schema
machine-readable description of a *repertoire* (3.8)

3.3
grapheme cluster
base character followed by zero or more continuing characters

Note 1 to entry: A grapheme cluster typically represents what the user thinks of as basic unit of a writing system for a language.

[SOURCE: UAX 29]

3.4
hull
set of code points or code point sequences (excluding high- or low-surrogate code points) that are not guaranteed to be excluded from the *repertoire* (3.8)

3.5
kernel
set of code points or code point sequences (excluding high- or low-surrogate code points) that are guaranteed to be included by the *repertoire* (3.8)

3.6
mode
option to specify whether characters or *grapheme clusters* (3.3) are examined

Note 1 to entry: The first edition did not have modes. Thus, characters can be examined, but grapheme clusters cannot.

3.7
registry
collection of named *repertoires* (3.8)

3.8
repertoire
description of a set of code points or code point sequences excluding high- or low-surrogate code points

4 Notation

in(x, A): code point or code point sequence x is in the repertoire described by a CREPDL element A;

not-in(x, A): code point or code point sequence x is not in the repertoire described by a CREPDL element A;

unknown(x, A): it is unknown whether code point or code point sequence x is in the repertoire described by a CREPDL element A.

NOTE 1 This predicate-like notation captures the combination of three-valued logic and the interpretation of a formula for a given character or grapheme cluster. In other words, in(x, A) implies that the interpretation of A under x is truth in three-valued logic. Likewise, not-in(x, A) and unknown(x, A) imply the interpretations of A under x are false and unknown, respectively.

NOTE 2 This document is intended to ensure that exactly one of in(x,A), not-in(x,A), and unknown(x,A) holds.

5 Overview

5.1 Basic constructs and compound constructs

Basic constructs of CREPDL schemas are created from regular expressions or references to registries of repertoires. Compound constructs of CREPDL schemas are created by combining basic constructs by set operators such as union, intersection, and difference.

5.2 Characters and code points

Although the title of this document is "Character Repertoire Description Language", this document uses code points more often than characters. This is because CREPDL allows the use of unassigned code points, which are not characters. For example, U+1CBB is an unassigned code point, and is thus not a character. It is possible to create a CREPDL schema that allows this code point. A stream containing it is valid against such a CREPDL schema.

5.3 Grapheme clusters

CREPDL can enable the validation of grapheme clusters, which are sequences of code points. For example, a CREPDL schema can allow LATIN CAPITAL LETTER N (U+004E) or LATIN SMALL LETTER n (U+006E) followed by COMBINING GRAVE ACCENT (U+0300) while disallowing other characters followed by COMBINING GRAVE ACCENT (U+0300). Likewise, a CREPDL schema can indicate which variation selector can follow which CJK unified ideograph.

NOTE The first edition cannot enable the validation of sequences of code points. It was thus not possible to allow LATIN CAPITAL LETTER N (U+004E) or LATIN SMALL LETTER n (U+006E) followed by COMBINING GRAVE ACCENT (U+0300) without allowing other characters followed by COMBINING GRAVE ACCENT (U+0300).

5.4 Kernel and Hull

It is sometimes difficult to precisely specify a repertoire. As an example, consider collections in ISO/IEC 10646, which are numbered and named repertoires. Some collections are open: they contain assigned code points as well as unassigned code points, which can be assigned in the future.

Recall that some basic constructs of CREPDL schemas are created from regular expressions. Such basic constructs have pairs of regular expressions. One regular expression specifies what is guaranteed to be included, while the other specifies what is not guaranteed to be excluded. The former and latter are called kernel and hull, respectively. If a code point matches the kernel regular expression, the code point is definitely included in the repertoire. Even if it isn't, it it not guranteed be excluded from the repertoire if it matches the hull regular expression.

NOTE Kernel and hull are reproduced from W3C Note-charcol[3]. Some examples in Annex B are also reproduced from W3C Note-charcol[3].

6 Syntax

6.1 General

A CREPDL schema shall be an XML document (which shall be as specified in W3C XML and shall further conform to W3C XML-Names) valid against the NVDL (ISO/IEC 19757-4) script in 6.3, which in turn relies on the RELAX NG (ISO/IEC 19757-2) schema in 6.2. The elements allowed in the RELAX NG schema

are of the namespace `http://purl.oclc.org/dsdl/crepdl/ns/structure/2.0`. Further constraints on the character content of the `char`, `kernel` or `hull` elements are shown in 6.4.

NOTE 1 W3C XML specifies that characters in XML documents are either U+0009 (CHARACTER TABULATION), U+000A (LINE FEED), U+000D (CARRIAGE RETURN), or a character in the ranges from U+0020 to U+D7FF, U+E000 to U+FFFD, or U+10000 to U+10FFFF. Since CREPDL schemas are represented by XML documents, other characters cannot directly occur in CREPDL schemas.

NOTE 2 The first edition used a different namespace name.

6.2 RELAX NG schema

```
# The following permission notice and disclaimer shall be included in
# all copies of this schema ("the Schema"), and derivations of
# the Schema:
#
# Permission is hereby granted, free of charge in perpetuity, to any
# person obtaining a copy of the Schema, to use, copy, modify, merge and
# distribute free of charge, copies of the Schema for the purposes of
# developing, implementing, installing and using software based on the
# Schema, and to permit persons to whom the Schema is furnished to do
# so, subject to the following conditions:
#
# THE SCHEMA IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR
# IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY,
# FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL
# THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR
# OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE,
# ARISING FROM, OUT OF OR IN CONNECTION WITH THE SCHEMA OR THE USE OR
# OTHER DEALINGS IN THE SCHEMA.
#
# In addition, any modified copy of the Schema shall include the following
# notice:
#
# THIS SCHEMA HAS BEEN MODIFIED FROM THE SCHEMA DEFINED IN ISO/IEC 19757-7,
# AND SHOULD NOT BE INTERPRETED AS COMPLYING WITH THAT STANDARD.

default namespace = "http://purl.oclc.org/dsdl/crepdl/ns/structure/2.0"

start = coll
coll =
  union | intersection | difference | ref | repertoire | char
union = element union { commonAtts, coll+ }
intersection = element intersection { commonAtts, coll+ }
difference = element difference { commonAtts, coll+ }
ref =
  element ref {
    commonAtts,
    attribute href { xsd:anyURI }
}
repertoire =
  element repertoire {
    commonAtts,
    attribute registry { text },
    attribute version { text }?,
     (attribute name { text } | attribute number {xsd:int})
}
char =
   element char {
     commonAtts,
      (text
       | element kernel { commonAtts, text }
      | element hull { commonAtts, text }
      | (element kernel { commonAtts, text },
         element hull { commonAtts, text }))
    }
commonAtts =
   attribute minUcsVersion { text }?,
   attribute maxUcsVersion { text }?,
```


