INTERNATIONAL STANDARD

ISO/IEC 19757-7

Second edition
2020-08

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 19757-7:2020(E)

© ISO/IEC 2020



COPYRIGHT PROTECTED DOCUMENT

© ISO/IEC 2020

All rights reserved. Unless otherwise specified, or required in the context of its implementation, no part of this publication may be reproduced or utilized otherwise in any form or by any means, electronic or mechanical, including photocopying, or posting on the internet or an intranet, without prior written permission. Permission can be requested from either ISO at the address below or ISO's member body in the country of the requester.

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

Published in Switzerland

Contents

Page

Foreword ... **iv**
Introduction ... **v**
1 Scope ... **1**
2 Normative references ... **1**
3 Terms and definitions ... **1**
4 Notation ... **2**
5 Overview ... **3**
5.1 Basic constructs and compound constructs ... 3
5.2 Characters and code points ... 3
5.3 Grapheme clusters ... 3
5.4 Kernel and Hull ... 3
6 Syntax ... **3**
6.1 General ... 3
6.2 RELAX NG schema ... 4
6.3 NVDL script ... 5
6.4 Regular Expressions ... 5
7 Semantics ... **5**
7.1 General ... 5
7.2 char ... 6
7.3 union ... 7
7.4 intersection ... 7
7.5 difference ... 7
7.6 ref ... 8
7.7 repertoire ... 8
8 Validation ... **8**
Annex A (informative) **Differences of conformant processors** ... **10**
Annex B (informative) **Example CREPDL schemas** ... **11**
Bibliography ... **15**

© ISO/IEC 2020 – All rights reserved

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.

© ISO/IEC 2020 – All rights reserved

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.

© ISO/IEC 2020 – All rights reserved

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/IEC 2020 – All rights reserved

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.

© ISO/IEC 2020 – All rights reserved