INTERNATIONAL STANDARD

ISO/IEC 19757-2

First edition
2003-12-01

Information technology — Document Schema Definition Language (DSDL) —

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

iTeh STANDARD PREVIEW

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

Partie 2: Validation de grammaire orientée courante — RELAX NG

(standards.iteh.ai)
ISO/IEC 19757-2:2003
https://standards.iteh.ai/catalog/standards/sist/bf40bae4-382d-40d4-a3b6-db8ffc51b5aa/iso-iec-19757-2-2003

Reference number
ISO/IEC 19757-2:2003(E)

© ISO/IEC 2003

PDF disclaimer

This PDF file may contain embedded typefaces. In accordance with Adobe's licensing policy, this file may be printed or viewed but shall not be edited unless the typefaces which are embedded are licensed to and installed on the computer performing the editing. In downloading this file, parties accept therein the responsibility of not infringing Adobe's licensing policy. The ISO Central Secretariat accepts no liability in this area.

Adobe is a trademark of Adobe Systems Incorporated.

Details of the software products used to create this PDF file can be found in the General Info relative to the file; the PDF-creation parameters were optimized for printing. Every care has been taken to ensure that the file is suitable for use by ISO member bodies. In the unlikely event that a problem relating to it is found, please inform the Central Secretariat at the address given below.

iTeh STANDARD PREVIEW
(standards.iteh.ai)

ISO/IEC 19757-2:2003
https://standards.iteh.ai/catalog/standards/sist/bf40bae4-382d-40d4-a3b6-db8ffc51b5aa/iso-iec-19757-2-2003

© ISO/IEC 2003

All rights reserved. Unless otherwise specified, no part of this publication may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying and microfilm, without permission in writing from either ISO at the address below or ISO's member body in the country of the requester.

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

 © ISO/IEC 2003 – All rights reserved

Contents

Page

iTeh STANDARD PREVIEW
(standards.iteh.ai)
ISO/IEC 19757-2:2003
https://standards.iteh.ai/catalog/standards/sist/bf10baa4-382d-40d4-a3b6-db88c51b5aa8/iso-iec-19757-2-2003

© ISO/IEC 2003 – All rights reserved

iTeh STANDARD PREVIEW
(standards.iteh.ai)
ISO/IEC 19757-2:2003
https://standards.iteh.ai/catalog/standards/sist/bf40bae4-382d-40d4-a3b6-db8ffc51b5aa/iso-iec-19757-2-2003

© ISO/IEC 2003 – 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. 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.

The main task of the joint technical committee is to prepare International Standards. Draft International Standards adopted by the joint technical committee are circulated to national bodies for voting. Publication as an International Standard requires approval by at least 75 % of the national bodies casting a vote.

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.

ISO/IEC 19757-2 was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 34, *Document description and processing languages*.

iTeh STANDARD PREVIEW
(standards.iteh.ai)

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

ISO/IEC 19757-2:2003
https://standards.iteh.ai/catalog/standards/sist/bf40bae4-382d-40d4-a3b6-db8ffc51b5aa/iso-iec-19757-2-2003

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

The following parts are under preparation.

— *Part 1: Overview*

— *Part 4: Selection of validation candidates*

Rule-based validation — Schematron, Datatypes, Path-based integrity constraints, Character repertoire validation, Declarative document manipulation, Datatype- and namespace-aware DTDs and Interoperability framework will form the subjects of future Parts 3, 5, 6, 7, 8, 9 and 10, respectively.

© ISO/IEC 2003 – All rights reserved

Introduction

The structure of this part of ISO/IEC 19757 is as follows. Clause 5 describes the data model, which is the abstraction of an XML document used throughout the rest of the document. Clause 6 describes the syntax of a RELAX NG schema. Clause 7 describes a sequence of transformations that are applied to simplify a RELAX NG schema, and also specifies additional requirements on a RELAX NG schema. Clause 8 describes the syntax that results from applying the transformations; this simple syntax is a subset of the full syntax. Clause 9 describes the semantics of a correct RELAX NG schema that uses the simple syntax; the semantics specify when an element is valid with respect to a RELAX NG schema. Clause 10 describes requirements that apply to a RELAX NG schema after it has been transformed into simple form. Finally, Clause 11 describes conformance requirements for RELAX NG validators.

This part of ISO/IEC 19757 is based on the RELAX NG Specification[1]. A tutorial for RELAX NG is available separately (see the RELAX NG Tutorial[2]).

iTeh STANDARD PREVIEW
(standards.iteh.ai)
ISO/IEC 19757-2:2003
https://standards.iteh.ai/catalog/standards/sist/bf40bae4-382d-40d4-a3b6-db8ffc51b5aa/iso-iec-19757-2-2003

© ISO/IEC 2003 – All rights reserved

Information technology — Document Schema Definition Language (DSDL) —

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

1 Scope

This part of ISO/IEC 19757 specifies RELAX NG, a schema language for XML. A RELAX NG schema specifies a pattern for the structure and content of an XML document. The pattern is specified by using a regular tree grammar. This part of ISO/IEC 19757 establishes requirements for RELAX NG schemas and specifies when an XML document matches the pattern specified by a RELAX NG schema.

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.

NOTE 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.

iTeh STANDARD PREVIEW
(standards.iteh.ai)

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

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

ISO/IEC 19757-2:2003
https://standards.iteh.ai/catalog/standards/sist/bf40bae4-382d-40d4-a3b6-db8ffc51b5aa/iso-iec-19757-2-2003

W3C XLink, *XML Linking Language (XLink) Version 1.0*, W3C Recommendation, 27 June 2001, available at <http://www.w3.org/TR/2001/REC-xlink-20010627/>

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

IETF RFC 2045, *Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies*, Internet Standards Track Specification, November 1996, available at <http://www.ietf.org/rfc/rfc2045.txt>

IETF RFC 2046, *Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types*, Internet Standards Track Specification, November 1996, available at <http://www.ietf.org/rfc/rfc2046.txt>

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

IETF RFC 2732, *Format for Literal IPv6 Addresses in URL's*, Internet Standards Track Specification, December 1999, available at <http://www.ietf.org/rfc/rfc2732.txt>

IETF RFC 3023, *XML Media Types*, Internet Standards Track Specification, August 1998, available at <http://www.ietf.org/rfc/rfc3023.txt>

3 Terms and definitions

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

3.1
resource
something with identity, potentially addressable by a URI

© ISO/IEC 2003 – All rights reserved

3.2
URI
compact string of characters that uses the syntax defined in IETF RFC 2396 to identify an abstract or physical resource

3.3
URI reference
URI or relative URI and optional fragment identifier

3.4
relative URI
form of URI reference that can be resolved with respect to a base URI to produce another URI

3.5
base URI
URI used to resolve relative URIs

3.6
fragment identifier
additional information in a URI reference used by a user agent after the retrieval action on a URI has been successfully performed

3.7
instance
XML document that is being validated with respect to a RELAX NG schema

iTeh STANDARD PREVIEW
(standards.iteh.ai)

3.8
space character
character with the code value #x20

ISO/IEC 19757-2:2003
https://standards.iteh.ai/catalog/standards/sist/bf40bae4-382d-40d4-a3b6-db8ffc51b5aa/iso-iec-19757-2-2003

3.9
whitespace character
character with the code value #x20, #x9, #xA or #xD

3.10
name
pair of a URI and a local name

3.11
namespace URI
URI that is part of a name

3.12
local name
NCName that is part of a name

3.13
NCName
string that matches the NCName production of W3C XML-Names

3.14
name class
part of a schema that can be matched against a name

3.15
pattern
part of a schema that can be matched against a set of attributes and a sequence of elements and strings

© ISO/IEC 2003 – All rights reserved

3.16
foreign attribute
attribute with a name whose namespace URI is neither the empty string nor the RELAX NG namespace URI

3.17
foreign element
an element with a name whose namespace URI is not the RELAX NG namespace URI

3.18
full syntax
syntax of a RELAX NG grammar before simplification

3.19
simple syntax
syntax of a RELAX NG grammar after simplification

3.20
simplification
transformation of a RELAX NG schema in the full syntax to a schema in the simple syntax

3.21
datatype library
mapping from local names to datatypes

NOTE a datatype library is identified by a URI

iTeh STANDARD PREVIEW
(standards.iteh.ai)

3.22
datatype
set of strings together with an equivalence relation on that set

ISO/IEC 19757-2:2003
https://standards.iteh.ai/catalog/standards/sist/bf40bae4-382d-40d4-a3b6-db8ffc51b5aa/iso-iec-19757-2-2003

3.23
axiom
proposition that is provable unconditionally

3.24
inference rule
rule consisting of one or more positive or negative antecendents and exactly one consequent, which makes the consequent provable if all the positive antecedents are provable and none of the negative antecendents is provable

3.25
valid with respect to a schema
member of the set of XML documents described by the schema

3.26
schema
specification of a set of XML documents

3.27
grammar
start pattern together with a mapping from NCNames to patterns

3.28
correct schema
schema that satisfies all the requirements of this part of ISO/IEC 19757

© ISO/IEC 2003 – All rights reserved

3.29
validator
software module that determine whether a schema is correct and whether an instance is valid with respect to a schema

3.30
path
list of NCNames separated by / or //

3.31
infoset
an abstraction of an XML document defined by W3C XML-Infoset

3.32
information item
constituent of an information set

3.33
data model
abstract representation of an XML document defined by this part of ISO/IEC 19757

3.34
XML document
string that is a well-formed XML document as defined in W3C XML

3.35
EBNF
Extended BNF
notation used to described context-free grammars

iTeh STANDARD PREVIEW
(standards.iteh.ai)
ISO/IEC 19757-2:2003
https://standards.iteh.ai/catalog/standards/sist/bf40bae4-382d-40d4-a3b6-db8ffc51b5aa/iso-iec-19757-2-2003

3.36
weak matching
kind of matching specified in detail in 9.3.7

3.37
in-scope grammar
nearest ancestor grammar element

3.38
content-type
one of the three values empty, complex, or simple

3.39
mixed sequence
sequence that may contain both elements and strings

4 Notation

4.1 EBNF

This part of ISO/IEC 19757 uses EBNF notation to describe the full syntax and the simple syntax of RELAX NG. A description of a grammar in EBNF consists of one or more production rules. Each production rule consists of the name of a non-terminal, followed by ::=, followed by a list of alternatives separated by |. Within an alternative, italic type is used to reference a non-terminal, concatenation indicates sequencing, [] indicates optionality, + indicates repetition one or more times and * indicates repetition zero or more times; other characters in normal type stand for themselves.

© ISO/IEC 2003 – All rights reserved

4.2 Inference rules

4.2.1 Variables

The symbol used for a variable indicates the variable's range as follows:

— *n* ranges over names

— *nc* ranges over name classes

— *ln* ranges over local names; a local name is a string that matches the NCName production of W3C XML-Names, that is, a name with no colons

— *u* ranges over URIs

— *cx* ranges over contexts (as defined in Clause 5)

— *a* ranges over sets of attributes; a set with a single member is considered the same as that member

— *m* ranges over sequences of elements and strings; a sequence with a single member is considered the same as that member; the sequences ranged over by *m* may contain consecutive strings and may contain strings that are empty

NOTE There are sequences ranged over by *m* that cannot occur as the children of an element.

— *p* ranges over patterns (elements matching the pattern production)

— *s* ranges over strings

— *ws* ranges over the empty sequence and strings that consist entirely of whitespace

— *params* ranges over sequences of parameters

— *e* ranges over elements

— *ct* ranges over content-types

iTeh STANDARD PREVIEW
(standards.iteh.ai)
ISO/IEC 19757-2:2003
https://standards.iteh.ai/catalog/standards/sist/bf40bae4-382d-40d4-a3b6-a0f1e91b5faa/iso-iec-19757-2-2003

4.2.2 Propositions

The following notation is used for propositions:

— *n* in *nc* means that name *n* is a member of name class *nc*

— $cx \vdash a$; *m* =~ *p* means that with respect to context *cx*, the attributes *a* and the sequence of elements and strings *m* matches the pattern *p*

— disjoint(a_1, a_2) means that there is no name that is the name of both an attribute in a_1 and of an attribute in a_2

— m_1 interleaves m_2; m_3 means that m_1 is an interleaving of m_2 and m_3

— $cx \vdash a$; $m =\sim_{weak} p$ means that with respect to context *cx*, the attributes *a* and the sequence of elements and strings *m* weakly matches the pattern *p*

— okAsChildren(*m*) means that the mixed sequence *m* can occur as the children of an element: it does not contain any member that is an empty string, nor does it contain two consecutive members that are both strings

© ISO/IEC 2003 – All rights reserved

— deref(*ln*) = <element> *nc p* </element> means that the grammar contains <define name="*ln*"> <element> *nc p* </element> </define>

— datatypeAllows(*u*, *ln*, *params*, *s*, *cx*) means that in the datatype library identified by URI *u*, the string *s* interpreted with context *cx* is a legal value of datatype *ln* with parameters *params*

— datatypeEqual(*u*, *ln*, s_1, cx_1, s_2, cx_2) means that in the datatype library identified by URI *u*, string s_1 interpreted with context cx_1 represents the same value of the datatype *ln* as the string s_2 interpreted in the context of cx_2

— $s_1 = s_2$ means that s_1 and s_2 are identical

— valid(*e*) means that the element *e* is valid with respect to the grammar

— start() = *p* means that the grammar contains <start> *p* </start>

— groupable(ct_1, ct_2) means that the content-types ct_1 and ct_2 are groupable

— $p :_c ct$ means that pattern *p* has content-type *ct*

— incorrectSchema() means that the schema is incorrect

4.2.3 Expressions

The following notation is used for expressions in propositions:

iTeh STANDARD PREVIEW
(standards.iteh.ai)

— name(*u*, *ln*) returns a name with URI *u* and local name *ln*

— m_1, m_2 returns the concatenation of the sequences m_1 and m_2

ISO/IEC 19757-2:2003
https://standards.iteh.ai/catalog/standards/sist/bf40bae4-382d-40d4-a3b6-db8ffc51b5aa/iso-iec-19757-2-2003

— $a_1 + a_2$ returns the union of a_1 and a_2

— ( ) returns an empty sequence

— { } returns an empty set

— "" returns an empty string

— attribute(*n*, *s*) returns an attribute with name *n* and value *s*

— element(*n*, *cx*, *a*, *m*) returns an element with name *n*, context *cx*, attributes *a* and mixed sequence *m* as children

— max(ct_1, ct_2) returns the maximum of ct_1 and ct_2 where the content-types in increasing order are empty( ), complex( ), simple( )

— normalizeWhiteSpace(*s*) returns the string *s*, with leading and trailing whitespace characters removed, and with each other maximal sequence of whitespace characters replaced by a single space character

— split(*s*) returns a sequence of strings one for each whitespace delimited token of *s*; each string in the returned sequence will be non-empty and will not contain any whitespace

— context(*u*, *cx*) returns a context which is the same as *cx* except that the default namespace is *u*; if *u* is the empty string, then there is no default namespace in the constructed context

— empty( ) returns the empty content-type

— complex( ) returns the complex content-type

© ISO/IEC 2003 – All rights reserved

— simple( ) returns the simple content-type
— [*cx*] within the start-tag of a pattern refers to the context of the pattern element

5 Data model

RELAX NG deals with XML documents representing both schemas and instances through an abstract data model. XML documents representing schemas and instances shall be well-formed in conformance with W3C XML and shall conform to the constraints of W3C XML-Names.

An XML document is represented by an element. An element consists of

— a name

— a context

— a set of attributes

— an ordered sequence of zero or more children; each child is either an element or a non-empty string; the sequence never contains two consecutive strings

A name consists of

— a string representing the namespace URI; the empty string has special significance, representing the absence of any namespace

— a string representing the local name; this string matches the NCName production of W3C XML-Names

iTeh STANDARD PREVIEW
(standards.iteh.ai)

A context consists of

— a base URI

ISO/IEC 19757-2:2003
https://standards.iteh.ai/catalog/standards/sist/bf40bae4-382d-40d4-a3b6-

— a namespace map; this maps prefixes to namespace URIs, and also may specify a default namespace URI (as declared by the xmlns attribute)

An attribute consists of

— a name

— a string representing the value

A string consists of a sequence of zero or more characters, where a character is as defined in W3C XML.

The element for an XML document is constructed from the infoset (see W3C XML-Infoset) of the XML document as follows. The notation [*x*] refers to the value of the *x* property of an information item. An element is constructed from a document information item by constructing an element from the [document element]. An element is constructed from an element information item by constructing the name from the [namespace name] and [local name], the context from the [base URI] and [in-scope namespaces], the attributes from the [attributes], and the children from the [children]. The attributes of an element are constructed from the unordered set of attribute information items by constructing an attribute for each attribute information item. The children of an element are constructed from the list of child information items first by removing information items other than element information items and character information items, and then by constructing an element for each element information item in the list and a string for each maximal sequence of character information items. An attribute is constructed from an attribute information item by constructing the name from the [namespace name] and [local name], and the value from the [normalized value]. When constructing the name of an element or attribute from the [namespace name] and [local name], if the [namespace name] property is not present, then the name is constructed from an empty string and the [local name]. A string is constructed from a sequence of character information items by constructing a character from the [character code] of each character information item.

© ISO/IEC 2003 – All rights reserved